



République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université des Sciences et de la Technologie Houari Boumediene

Faculté d'Electronique et d'Informatique
Département Informatique

Mémoire de Master

Option
Ingénierie de logiciel

Thème

Extraction d'événements à partir de textes écrits
en langue arabe

Encadreur :

Dr. ALIANE H., CERIST

Présenté par :

GUENDOUZI Wassila.
MOKRANI Amina.

Devant le jury composé de :

Pr. A. Guessoum	USTHB	Président
Mme. M. Mameri	USTHB	Membre
Mme Benchouk	USTHB	Membre

Binôme N°/0542012

Remerciements

Tout d'abord, nous remercions *DIEU* le tout puissant de nous avoir aidées pour accomplir ce modeste travail.

Nous tenons à remercier les membres de jury : Pr. Guessoum, Mme Mameri et Mme Benchouk de nous avoir fait l'honneur de juger ce travail.

Un Grand Merci et une reconnaissance sincère à notre promotrice *Mme ALIANE* pour nous avoir fait découvrir le domaine passionnant du TAL ainsi que pour sa disponibilité, sa bienveillance et son aide fructueuse qui nous a permis de mener à bien ce travail.

Une reconnaissance particulière à toutes les personnes qui ont contribué de loin ou de prêt à l'aboutissement de ce travail, pour leurs encouragements, suggestions et remarques ; en gardant toujours un grand sourire.

Enfin, nous ne pouvons qu'être infiniment reconnaissantes envers nos parents pour leur soutien indescriptible, leur patience, leur confiance et leurs nombreux sacrifices. Nous leur dédions avec plaisir ce travail Qu'ils sachent que nous sommes conscientes de ce que nous leur devons et qu'ils trouvent ici l'immense témoignage de notre affection

Résumé

Avec l'augmentation sans cesse grandissante des volumes de l'information textuelle sur le web, le problème de l'extraction de connaissances à partir de données textuelles devient crucial et incontournable. Le processus d'extraction de connaissances inclut généralement plusieurs étapes, dont la reconnaissance d'entités et d'événements dans lesquelles elles sont impliquées, ainsi que les informations temporelles, spatiales et autres qui peuvent leur être rattachées. Néanmoins, bien que le problème de l'identification des entités nommées soit relativement bien compris, l'identification d'événements et de leurs relations temporelles demeure un problème qui est encore peu abordé. Or, que ce soit pour les applications de réponses à des questions, ou la simple recherche d'information, on est souvent appelé à rechercher ou annoter dans les documents les éléments qui nous permettent de répondre à des questions de la forme suivante : Quand un certain événement a-t-il eu lieu ? Que s'est-il passé pendant une certaine période de temps donnée ? Dans quels événements une personne a-t-elle été impliquée ?

Les événements peuvent être désignés par des verbes mais on peut aussi avoir des événements nominaux désignés comme des déverbaux. Une étape du travail d'extraction des événements est de déterminer les expressions désignant potentiellement un événement, puis de déterminer en contexte si c'est bien le cas.

Nous proposons dans ce travail une approche pour l'extraction des entités linguistiques désignant potentiellement des événements ainsi que les expressions pouvant y afférer à savoir les expressions temporelles et les expressions de lieu. Nous n'utilisons pas de corpus annoté ni de lexiques mais notre approche est fondée sur l'élaboration d'un ensemble de règles qui utilisent des indicateurs formels issus de l'observation du corpus de travail. Nous travaillons sur un corpus de dépêches de presse écrites en Arabe Standard Moderne, et les résultats que nous obtenons sont comparables aux résultats des travaux que nous avons étudiés pour les autres langues.

Mots clés :

TALN, langue arabe, segmentation des mots arabes, annotation en partie de discours, extraction de connaissance, fouille de texte, extraction d'événement, expressions temporelles expressions de lieu..

Liste des tableaux

Tableau 4.1	liste des marqueurs temporels autonomes	53
Tableau 4.2	liste des marqueurs temporels déclencheurs	54
Tableau 4.3	liste des marqueurs temporels délimiteurs	54
Tableau 4.4	liste des marqueurs de lieu (déclencheur et autonomes)	59
Tableau 4.5	Evaluation du résultat de l'extraction d'événement	62

Liste des figures

Figure 1.1	Schéma générale d'un système de dialogue	6
Figure 1.2	Schéma générale d'un système de traduction	6
Figure 1.3	Étapes de l'analyse	7
Figure 1.4	Schéma global de l'ECBD	13
Figure 1.5	Schéma général du processus de fouille de texte	17
Figure 1.6	Schéma illustrant les types d'entités nommées	21
Figure 2.1	Exemple de l'analyse syntaxique	34
Figure 4.1	Schéma illustrant la modélisation d'événement	45
Figure 4.2	Schéma général de l'approche	46
Figure 4.3	Schéma générale de l'algorithme de segmentation	48
Figure 4.4	liste des ajouts nominaux (à gauche et à droite)	49
Figure 4.5	liste des ajouts à droite des verbes inaccomplis futurs et inaccomplis présents	50
Figure 4.6	liste des ajouts à gauche des verbes accomplis	51

Figure 4.7	exemple sur l'extraction d'expression temporelle par la règle R1	56
Figure 4.8	exemple 1 sur l'extraction d'expression temporelle par la règle R2	56
Figure 4.9	exemple 2 sur l'extraction d'expression temporelle par la règle	57
Figure 4.10	schéma illustrant le principe général d'extraction des expressions temporelles	58
Figure 4.11	exemple sur l'extraction de lieu par la règle R3	56
Figure 4.12	Schéma illustrant le principe général d'extraction des lieux	60
Figure 5.1	l'interface graphique du système	66
Figure 5.2	la barre de tâche	67
Figure 5.3	résultat de bouton « segmentation »	68
Figure 5.4	résultat de bouton « annotation partielle en POS »	69
Figure 5.5	résultat de bouton « annoter infos événement »	70

Table de matières

<u>INTRODUCTION GENERALE</u>	1
<u>CHAPITRE 1 : TALN et Extraction de Connaissances</u>	3
1. Introduction	3
2. Traitement automatique du langage naturel	3
2.1. Aperçus historique	3
2.2. A quoi sert le TALN	5
2.3. Qu'est ce que le TALN	5
2.3.1. Analyse et génération	5
2.3.2. Niveau linguistique	7
2.4. Les applications du TALN.....	9
2.4.1. Le traitement documentaire.....	9
2.4.2. La production de document	10
2.4.3. Les interfaces en langage naturel.....	10
3. Extraction de connaissances.....	11
3.1 De l'information à la connaissance	11
3.2 Extraction d'information et extraction de connaissances	11
3.3 Extraction de connaissance à partir des BD	12
3.4 Extraction de connaissances à partir des textes	14
3.4.1. Aperçus historique	15
3.4.2. Définition de la fouille de textes.....	15
3.4.3. Fouille de textes et fouille de données	16
3.4.4. Processus de fouille de textes	17
3.4.4.1. Préparation du corpus.....	17
3.4.4.2. Extraction terminologique.....	18
3.4.4.3. Détection des traces de concept.....	18
3.4.4.4. Extraction de connaissances	18
3.4.5. TALN et l'extraction de la connaissance à partir du texte.....	19

4. Extraction d'entités nommées	20
4.1. Qu'est ce que l'entité nommée.....	20
4.2. Les différents types d'entités nommées.....	20
4.3. Reconnaissance des entités nommées	21
4.4. Applications	22
4.4.1. Extraction des informations du texte.....	22
4.4.2. Répondre automatiquement à des questions.....	22
4.4.3. Améliorer les résultats des systèmes de recherche	22
5. Conclusion	22
<u>CHAPITRE 2 : Extraction d'Événements</u>	23
1. Introduction	23
2. Concept d'événement	23
3. Concept d'expression temporelle	24
4. Extraction d'informations temporelles	24
5. Types d'approches d'extraction d'événement	25
5.1. Approches symboliques	26
5.2. Approches d'apprentissage	26
5.2.1 Méthodes supervisées.....	26
5.2.2 Méthodes semi-supervisées	27
5.2.3 Méthodes non supervisées	27
5.3. Approches Hybrides	27
6. Méthodes d'extraction d'événements existantes	28
6.1. Méthode suivie par le système EVITA	28
6.2. La stratégie de l'exploration contextuelle.....	30
6.3. Méthode basée sur la chaîne lexicale	32
6.4. Méthode basée sur un analyseur syntaxique	34
7. Conclusion	35
<u>CHAPITRE 3 : Langue Arabe</u>	36
1. Introduction	36
2. Particularités de la langue arabe	36
3. Morphologie arabe.....	37

4. Problèmes du traitement automatique de l'arabe	38
4.1. Problème de voyellation	38
4.2. Structure complexe des mots	39
4.3. Flexibilité de l'ordre des mots dans une phrase simple	39
4.4. Ambiguïté lexicale et syntaxique	40
5. Conclusion	40
<u>CHAPITRE 4 : Une Approche Linguistique pour l'Extraction d'Evénements à partir de Textes en Langue Arabe</u>	41
1. Introduction	41
2. Langue cible	41
3. Retour sur les travaux reliés	43
4. Conception d'événement	44
5. Approche adoptés	45
5.1. La segmentation	47
5.2. Annotation partielle en parties de discours	48
5.3. Annotation des événements verbaux	52
5.4. Annotation des expressions temporelles	53
5.4.1. Identification des marqueurs temporelles.....	53
5.4.2. Règles de l'analyse contextuelle.....	55
5.4.3. Algorithme générale de l'annotation des expressions temporelles.....	57
5.5. Annotation des expressions de lieu	58
5.5.1. Identification des marqueurs de lieu	58
5.5.2. Règles de l'analyse contextuelle.....	59
5.5.3. Algorithme générale de l'annotation des expressions de lieu	60
6. Résultat et évaluation de l'extraction d'événement	60
7. Conclusion	63
<u>CHAPITRE 5 : Mise en œuvre</u>	64
1. Introduction	64
2. Langage de développement.....	64
2.2. Pourquoi choisir python.....	64
2.3 Domaine d'application du python	64

3. Environnement de développement	65
4. Fonctionnalités du système	65
4.1. Zone de texte.....	67
4.2. Barre de tâche.....	67
4.3. Boutons de traitement.....	67
4.3.1 Boutons de segmentation.....	68
4.3.2 Boutons d'annotation partielle en POS	69
4.3.3 Boutons extraction d'événement.....	70
<u>CONCLUSION GENERALE ET PERSPECTIVES</u>	71
<u>REFERENCES BIBLIOGRAPHIQUES</u>	73
<u>ANNEXE A : PYTHON</u>	77
<u>ANNEXE B : RESULTAT ET EVALUATION</u>	86