REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE MOHAMED BOUDIAF - M'SILA
FACULTE DES MATHEMATIQUES ET DE L'INFORMATIQUE

DEPARTEMENT D'INFORMATIQUE

MEMOIRE de fin d'études

Présenté pour l'obtention du diplôme de MASTER

Domaine : Mathématiques et Informatique

Filière : Informatique

Spécialité : Systèmes d'Informations Avancés

Par: NOUR Fatima Zahra

SUJET

Etude et implémentation d'une méthode morphologique pour l'extraction des racines des mots arabes

Soutenu publiquement le : 31 / 05 /2016 devant le jury composé de :

.....	**Université de M'sila**	**Président**
MAHDJOUBI Roussafi	**Université de M'sila**	**Rapporteur**
.....	**Université de M'sila**	**Examinateur**
.....	**Université de M'sila**	**Examinateur**

Promotion : 2015 /2016

TABLE DES MATIERES

Le traitement automatique du langage naturel est une discipline très importante qui ne cesse, depuis plusieurs années, de se développer et de gagner de plus en plus de terrain dans le domaine des différentes langues à travers le monde, bien que ce domaine soit très difficile à maîtriser étant donné que le langage humain est généralement très complexe et ne dispose d'aucunes règles fixes qui s'appliquent à la totalité des aspects de même nature. La majorité des travaux connus menés jusqu'ici concernaient les langues occidentales telles que l'anglais et le français. Le traitement de la langue arabe n'a vu le jour que quelques années plus tard après les premiers travaux. En effet, plusieurs tentatives dès lors, ont été menées, donnant des résultats plus ou moins satisfaisants. Ces tentatives ont clairement révélé l'aspect morphologique très complexe de cette langue, qui, à la différence des autres langues, possède une structure et des caractéristiques très spécifiques qui la rendent très particulière et augmentent ainsi considérablement l'effort nécessaire pour la maîtriser. Les techniques utilisées pour traiter la morphologie sont donc très diverses et variées, certaines sont inspirées des travaux menés sur les langues étrangères et d'autres reposent sur les propres caractéristiques de cette langue. Dans ce mémoire, nous allons exposer une contribution de notre part qui consiste à essayer d'appliquer l'une de ces méthodes et étudier la faisabilité d'un tel traitement après bien sûr, l'exposition d'une bonne partie portant sur l'étude des différents aspects morphologiques de cette langue nécessaires pour la suite du travail. D'une façon plus claire, le mémoire sera organisé de la manière suivante :

Dans le premier chapitre intitulé "**La langue arabe** ", nous allons donner un aperçu sur la langue arabe et exposer les différents aspects de sa morphologie ainsi que ses caractéristiques.

Puis, au sein de "**l'analyse morphologique de la langue arabe**", deuxième chapitre de ce travail, nous allons présenter le processus général d'extraction de la racine d'un mot arabe et les différentes techniques utilisées pour ce faire.

Finalement, le troisième et dernier chapitre qu'on a nommé "**conception et réalisation**", servira à exposer une conception de la méthode choisie, suivie de sa réalisation avec une présentation de quelques unes de ses interfaces.

CONCLUSION GENERALE

L'utilisation de la langue arabe comme moyen de communication à travers le support informatique a été longtemps appréhendée avec beaucoup d'hésitation par la communauté scientifique, notamment celle du monde arabe où cet outil trouvera beaucoup d'utilisations importantes. En effet, la langue et les différentes difficultés qui s'y rattachent, notamment le problème de l'ambiguïté issue de l'absence des voyelles, le problème de reconnaissance des formes fléchies (la langue arabe étant fortement flexionnelle) et le problème du manque de diversité des sujets traitant le domaine du traitement morphologique de la langue arabe se limitant à juste une partie de ce dernier, tout cela pose un énorme défi difficile à surmonter.

Malgré tout cela et malgré la courte durée consacrée à la réalisation de ce sujet, nous avons osé nous aventurer dans ce domaine et on peut dire que, vu les résultats obtenus, nous pensons qu'on a quand même pu relever ce défi et par la même occasion apprendre beaucoup de nouvelles connaissances tout au long de la réalisation de ce travail telles que la programmation objet (java) et les concepts du traitement automatique du langage.

Toutefois, le sujet étant très vaste, il reste beaucoup à faire pour améliorer ce travail, on peut donc proposer comme perspectives, l'introduction d'un plus grand nombre de racines et de schèmes ainsi que l'extension du traitement pour prendre en compte les mots défectueux.

Bibliographie :

[1] A. Chen, et F. Gey,"Building an Arabic stemmer for information retrieval" .TREC 2002. Gaithersburg: NIST, pp 631-639, 2002.

[2] A. Goweder, M. Poesio, A. De Roeck et J. Reynolds, "Identifying Broken Plurals in Unvowelised Arabic Text". Proceedings of EMNLP, pp. 246-253, 2003.

[3] Abd El Salam AL HAJJAR, "extraction et gestion de l'information a partir des documents arabes", thèse de Doctorat, Université Paris VIII, Saint Denis ,2010.

[4] Abd El Salam AL HAJJAR, Mohammad HAJJAR, Khaldoun ZREIK," Classification of Arabic Information Extraction methods",2nd International Conference on Arabic Language Resources and Tools Cairo (Egypt), pp. 22 – 23, 2009.

[5] Abd El Salam AL HAJJAR, Mohammad HAJJAR, Khaldoun ZREIK, "A system for evaluation of arabic root extraction methods", Fifth International Conference on Internet and Web Applications and Services, Barcelone, 2010.

[6] Aïda KHEMAKHEM, "ArabicLDB : une base lexicale normalisée pour la langue arabe", Mémoire de master, université de Sfax, Tunisie, 2006.

[7] Al Ajeeb Al Ajeeb, Sakher Company, website: http://lexicons.ajeeb.com. 2010.

[8] A. N. De Roeck, et W. Al-Fares, "A morphologically sensitive clustering algorithm for identifying Arabic roots". Proceedings ACL-2000. HongKong, pp. 199 – 206, 2000.

[9] Atef Ben Youssef, "Méthodes Mixtes pour la Traduction Automatique Statistique", Mémoire de master, université Stendhal, Grenoble3, 2008.

[10] B. Hammo, H. Abu-Salem, S. Lytinen, et M. Evens , "A Question Answering System to Support the Arabic Language ", Proceedings of the ACL-02 workshop on Computational approaches to Semitic languages Philadelphia, Pennsylvania, pp. 1 – 11, 2002.

[11] Ben Taamallah Sahnoun, "Prétraitement de données et création d'un segmenteur de l'arabe pour un système de traduction probabiliste vers le français", Mémoire de master, université Stendhal, Grenoble3, 2012.

[12] Boulaknadel Siham "Traitement Automatique des Langues et Recherche d'Information en langue arabe dans un domaine de spécialité: Apport des connaissances morphologiques et syntaxiques pour l'indexation", thèse de doctorat, Université de Nantes, 2008.

[13] CHAAR France," Traitement automatique de la langue arabe : La modération en arabe ", Mémoire de master, Institut National des Langues et Civilisations Orientales, Paris, 2012.

[14] Dilekh Tahar, "Implémentation d'un outil d'indexation et de recherche des textes en arabe", Mémoire de Magister, université Hadj Lakhdar , Batna,2011.

[15] Ed-dariouache Adnane, "Etude et réalisation d'un analyseur morphologique de la langue arabe", mémoire de master, UNIVERSITE Sidi Mohamed Ben Abdellah, Fès, 2015.

[16] F. Abu Hawas, "Exploit relations between the word letters and their placement in the word for arabic root extraction ", computer Science, vol. 14, no. 2, pp. 327-341, 2013.

[17] Farag Ahmed et Andreas Nürnberger," N-Grams Conflation Approach for Arabic Text", Proceedings of the International Workshop on improving Non English Web Searching(iNEWS07) In conjunction with The 30th Annual International (ACM SIGIR) Conference. Amsterdam City, Netherlands, pp. 39-46, 2007.

[18] Fouad Soufiane Douzidia ," Résumé automatique de texte arabe", Mémoire de M.Sc, Université de Montréal,2004.

[19] G. A. Kiraz, "Analysis of the Arabic Broken Plural and Diminutive", Proceedings of the 5th International Conference and Exhibition on Multi-Lingual Computing (ICEMCO96), Cambridge, UK, 1996.

[20] H. Al Ameed, S. Al Ketbi, A. Al Kaabi, K. Al Shebli, N. Al Shamsi, N. Al Nuaimi, et S. Al Muhairi, "Arabic Light Stemmer: A new Enhanced Approach" , The Second International Conference on Innovations in Information Technology (IIT'05), 2005.

[21] H. Wehr," Dictionary of Modern Written Arabic". Publié par Harrassowitz -Germany, 1961.

[22] K. Darwish," Al-stem: A light Arabic stemmer, 2002". Available: http : //www.glue.umd.edu/~kareem/research.

[23] K. R. Beesley, "Finite-State Morphological Analysis and Generation of Arabic at Xerox Research: Status and Plans in 2001". The ACL 2001Workshop on Arabic Language Processing: Status and Prospects, Toulouse, France, 2001.

[24] K. Taghva, R. Elkoury , et J. Coombs, "Arabic Stemming without a root dictionary", International Conference on Information Technology: Coding and Computing (ITCC'05) - Volume I pp. 152-157, 2005.

[25] L. Khreisat, "Arabic Text Classification Using N-gram Frequency Statistics a Comparative Study". The 2006 International Conference on Data Mining Part of the 2006 World Congress in Computer Sciences DMIN, pp. 78-82, 2006.

[26] L. Larkey, L. Ballesteros, et M. Connell," Light Stemming for Arabic IR Arabic Computational Morphology: Knowledge-based and Empirical Methods", A.Soudi, A. van en Bosch, and Neumann, G., Editors. Kluwer/Springer's serieson Text, Speech, and Language Technology, 2005.

[27] M. Aljlayl et Frieder Aljlayl, et O. Frieder," On arabic search: Improving the retrieval effectiveness via a light stemming approach". Proceedings of ACM Eleventh Conference on Information and Knowledge Management, Mclean, VA, pp. 340 - 347 , 2002.

[28] M. Mustafa, H. AbdAlla, et H. Suleman, "Current Approaches in Arabic IR: A Survey". Proceedings The Annual International Conference on Asia-Pacific Digital Libraries (ICADL), Bali, Indonesia. 2008.

[29] M. Rashwan, M. Al-Badrashiny, M. Attia, S.M Abdou, "A hybrid system for automatic arabic diacritization", The 2nd International Conference on Arabic Language Resources and Tools, Egypt, 2009.

[30] M. Sanan," Etude Des Méthodes De La Recherche D'information Et De L'indexation Sur Les Documents Electroniques : Cas De La Langue Arabe", Thèse de Doctorat, UNIVERSITE PARIS VIII - SAINT DENIS, 2008.

[31] Maraoui Mohsen , Zrigui Mounir, Antoniadis Georges," Un système de génération automatique de dictionnaires étiquetés de l'arabe",CITALA 2007, Rabat, Maroc

[32] N. Yousef, A. Abu-Errub, A. Odeh, et H. Khafajeh, ," An improved arabic word's roots extraction method using n-gram technique", Journal of Computer Science 10, pp. 716-719, 2014.

[33] R. Al Shalabi, et N. Evens, "A Computational Morphology System for Arabic", Proceedings of COLING-ACL, New Brunswick, NJ, pp. 66-72, 1998.

[34] R. Blachère, M. Gaudefroy-Demombynes, "Grammaire de l'arabe classique", Edition Maisonneuvre-Larose, Paris, 1975.

[35] S. Al-Fedaghi et H. Al-Sadoun, "Morphological compression of arabic text," in Information Processing & Management, pp. 303-316, 1990.

[36] S. Khoja and R. Garside, "Stemming Arabic text". Computing Department, Lancaster University,Lancaster, www.comp.lancs.ac.uk/computing/users/khoja/stemmer.ps, 1999.

[37] S. Mesfar,"Analyse morpho-syntaxique automatique et reconnaissance des entités nommées en arabe standard". Thèse de doctorat, université de Franche-Comté, 2008.

[38] Y. Kadri, A. Benyamina, "Système d'analyse syntaxicosémantique du langage arabe", *mémoire d'ingénieur, université d'Oran Es-sénia, 1992*

ملخص:

العمل المنجز في اطار هذه المذكرة يتعلق بتصميم برنامج يتيح استخراج جذور الكلمات العربية والذي هو جزء من المعالجة الآلية للغة. لهذا الغرض تم في بداية هذا العمل عرض دراسة عن بنية اللغة العربية متبوعا بعرض لمختلف التقنيات المقترحة في هذا المجال.

الكلمات المفتاحية : المعالجة الآلية للغة ، اللغة العربية، الجذر، الوزن، اللواحق، التجذيع، قاموس، التحليل البنيوي.

Abstract :

The work undertaken as part of this memory relates to the production of a system for extracting the roots of Arabic words that is part of natural language processing. For this, a study of Arabic morphology and a presentation of the different techniques performed in this area were initially exposed.

Key words : NLP, arabic language, root, scheme, affixes, lemmatisation, corpus, morphological analysis.

Résumé

Le travail entrepris dans le cadre de ce mémoire concerne la réalisation d'un système d'extraction des racines des mots arabes qui entre dans le cadre du traitement automatique du langage naturel. Pour ce faire, une étude sur la morphologie arabe et une présentation des différentes techniques réalisées dans ce domaine ont été initialement exposées.

Mots clés : TALN, langue arabe, racine, schème, affixes, lemmatisation, corpus, analyse morphologique.