Al Jinan الجنان

Volume 10 Article 22

2018

Extraction d'information à partir des sites Web en arabe basée sur une méthode à base des règles

Moustafa Alhajj
Lebanese University, moustafa.alhajj@gmail.com

Amani Sabra
Lebanese University, amani.sabra@hotmail.fr

Moustafa Al-Hajj / Amani Sabra
Centre des Sciences du Langage et de la Communication
Faculté des Lettres de l'Université Libanaise, Tayouneh, centre Céline – Liban

Extraction d'information à partir des sites Web en arabe basée sur une méthode à base des règles

DOI: 10.33986/0522-000-010-015

Résumé: Cet article décrit un outil qui se sert de l'ingénierie de la langue pour l'extraction d'information à partir des sites web en arabe, Ces informations serviront aux documentalistes du Web poue créer des fiches d'archivage pour les sites. Une fiche d'archivage est proposée, l'objectif étant de remplir cette fiche automatiquement. Pour la reconnaissance et la classification des segments textuels, la méthode d'exploration contextuelle proposée par Descles est utilisée, les marqueurs et règles linguistiques sont définis en se basant sur une étude synthétique des spécificités de la langue arabe. Un corpus de plus de 1300 sites Web en langue arabe a été construit, les résultats obtenus montrent l'intérêt de l'approche pour constituer des fiches d'archivage des sites Web en langue arabe(1).

Mots Clés: Exploration Contextuelle, Fouille de données textuelles, Extraction d'information, XML.

Introduction

L'archivage de sites Web s'inscrit dans un domaine plus large de l'archivage de document en général. Dans une perspective de conservation du Web, plusieurs tentatives d'archivage ont été faites, parmi lesquelles les projets de «Internet Archive»(2), une organisation lancée depuis 1996, consacrée à l'archivage du Web et qui a pour principale activité la prise des captures de toutes les pages des sites Web à intervalles régulières (chaque deux mois). Depuis août 2006,

(1) Cet article est le fruit d'un travail réalisé dans le cadre du projet de recherche "L'ingénierie de langue pour l'archivage automatique des sites Web en langue arabe" achevé en mai 2013.

(2) http://archive.bibalex.org/

la Bibliothèque nationale de France a pour mission de collecter et conserver les sites Internet du «domaine français» au titre du dépôt légal, et ce au moins une fois par an (à l'exception des sites comme ceux des journaux en ligne qui font l'objet d'une collecte quotidienne) , cependant l'accès aux ressources de la «Bibliothèque de recherche» reste restreint [MASANES, 2001]. La BAC (Bibliothèque et Archives Canada) a aussi été parmi les premiers à envisager l'archivage du Web(1). Le projet EVA,Ccoordonné par « Helsinki University Library »,Aest également un projet ayana pour principal objectif de tester des méthodes de capture, d'enregistrement, et de préservation de l'ensemble du Web finlandais [Lounamaa & Salonharju, 1999]. Le site http://www.webarchivists.org(2) fait le point sur l'état des travaux dans le domaine.

A l'issue de l'archivage des sites Web, plusieurs questions peuvent être soulevées sur les frontières d'un site, en particulier celles relatives au temps et au contenu;: quand est-ce qu'on peut faire l'archivage du site, et que choisir du site pour l'archive?

Les documentalistes ou experts qui s'occupent de l'archivage des sites Web, étant face à une gigantesque taille d'information sur le Web, auront besoin d'outils d'aide tels que des outils de synthèse, des systèmes de filtrage d'informations, des plates-formes de résumé automatique, des outils d'indexation et des outils d'aide à l'extraction des mét données sur les documents.

Dans la littérature, il existe des travaux qui abordent la question d'extraction d'information à partir de documents Web en langue arabe ([AL-Smadi & Qawasmeh, 2016] ; [Feriel & Khireddine, 2015] ; [Elsayed & Elghazaly, 2015]; [Alruily & Alghamdi, 2015] ; [Shaalan, 2014] ; [Al-Hajj & Mourad]), mais rares sont les travaux ([Mourad, 2002]) qui traitent la question de l'extraction d'information en vue de l'archivage du Web en arabe. Le présent travail se veut une contribution dans ce domaine , un outil basé sur une méthode à base des règles utilisant des marqueurs linguistiques en arabe a ainsi été mis en place. Cet outil s'appuie sur des techniques de détection de structures des pages Web, de filtrage, de marquage sémantique, et plus généralement sur des techniques de fouille de texte. L'accent est mis sur la robustesse, la performance, et le traitement de larges volumes de données.

(1) http://www.collectionscanada.gc.ca/webarchives/index-e.html

(2) Visité le 30 - 04 - 2016

Dans cet article, au §2 la méthode adoptée pour la construction du corpus est présentée. Au §3 les informations d'archivage pour les sites sont décrites avant de présenter la méthodologie au §4. Enfin, une discussion des résultats est exposée au §5.

Constitution du corpus

Pour définir les éléments nécessaires à l'archivage des sites et les règles linguistiques pour leur extraction, il est tout d'abord indispensable de construire un corpus. Le corpus servira à la fois d'étude et de test. L'étude est faite sur une partie des sites du corpu, ; elle vise à définir les pages pertinentes des sites pour les soumettre au traitement,r ainsi que les marqueurs linguistiques et les règles adaptés à la description des sites. Le test (ou le traitement) consiste à appliquer l'outil développé sur l'autre partie du corpus afin de vérifier la validité de l'approche pour l'archivage des sites Web en arabe.

Le choix des sites Web du corpus porte sur des critères variés tels que la variété des auteurs, des serveurs d'hébergement ainsi que des domaines d'hébergement (« lb » pour Liban, « ma » pour Maroc, « sy » pour Syrie, etc.) et sur le degré de popularité des sites.

Pour sélectionner les sites sur internet, une recherche avancée a été menée sur le moteur de recherche Google(1). Les mots cléssutilisés sont tels que وزارة, صحيفة, مشروع, سياسة , جامعة, مؤسسة, فنان …, (ministère, projet, politique, université, entreprise, artiste, …) et le critère de recherche choisi est « page en arabe ». A partir des 10 premiers résultats de ces différentes recherches, uniquement les pages d'accueils sont gardées. Le corpus comporte ainsi plus de 1300 sites Web (l'Annexe Isà la fin de l'article représente un échantillon de 275 sites Web du corpus), dont des sites ministériels, institutionnels, des organisations gouvernementales, des journaux, des universités, etc.

Information d'archivage pour les sites

Une étude préalable [Mourad, 2002] surdles informations utiles pour l'archivage dans les sites Web du corpus a permis de ressortir des informations suivantes :

Une fiche contenant l'identification du site etsd'informations le concernant, comme le nombre de pages et leul'organisatios, la typologie du site, sa notoriété, les sites associés, ainsi qu'une description sommaire de son contenu

(1) http://www.google.com

au niveau terminologique et des formats des fichiers.

Une extraction des séquences saillantes, pertinentes, représentatives du site.

Un résumé textuel du contenu du site.

A partir de ces constats, et d'après une étude des informations intéressantes pour l'archivage à partir des sites du corpus, la fiche d'archivage proposée dans Mourad, [2002] a été adaptée et étendue pour avoir la structure suivante:

Etiquette	Exemples	Commentaire
Adresse du site (عنوان الموقع)	http://www.a-alamri.com http://www.islamspirit.com http://alfeker.net/ http://aljubailtoday.com	L'adresse url du site
Date de l'archivage (تاريخ الارشفة)	26-04-2012	La date du traitement du site par l'outil
Titre du site (اسم الموقع)	مفكرة الاسلام (http://www.islammemo.cc/)	L'intitulé du site
Description (الوصف)	موقع اخباري متخصص عن مدينة نابلس والضفة ويقدم تغطية شاملة ومتواصلة (http://www.nablustv.net)	Une description du site décrivant son activité, etc.
Présentation (العرض)	وهذا الموقع هو كلمه أردت بواسطتها أن أتغلب على صمتي الشخصي (http://www.alameldin.net)	Une présentation du site
Objectif (الهدف)	يهدف الموقع أساسا إلى تزويد طلاب مرحلتي البكالوريوس والماجستير والباحثين بالمقررات الدراسية والأبحاث العلمية والمشاريع البحثية في تخصص الجيوفيزياء ... (Site : http://www.a-alamri.com)	L'objectif du site
Explication (شرح الموقع)	هذا الموقع يشرح الكاميرة المذدوجة في ال٧بلاس و خصائصها من زووم لتصوير بوكيه او بورتريت و معالج الصور الخارق الجديد ملحوظة معالج الصور الجديد يستخدم جزء من ... (https://www.electrony.net)	Une explication sur le site à l'aide d'une synthèse par exemple
Offre du site (تقديمات الموقع)	هذا الموقع يقدم عدة خدمات، منها: - قسم موسوعات: حيث يمكن تحميل أمهات الكتب الإسلامية في إصدارات موسوعية، تضم الكتب والمؤلفات التي تم... (http://www.islamspirit.com)	Les offres du site pour les utilisateurs
Personnes (الاشخاص المسؤولين)		Les noms des personnes responsables du site, des webmasters, ...
Mails (عناوين الكترونية)		Les adresses e-mails des personnes à contacter
Copyright (حقوق النشر)		Les informations copyright du site

Fiche d'archivage pour les sites Web

Méthodologie

Pour l'extraction d'information, la méthode d'exploration contextuelle proposée par Descles [DESCLES, 97] a été utilisé . Il s'agit d une méthode à

base des règles qui utilise des indices linguistiques de surface et qui permet une analyse sémantique des textes et l'identification des segments textuels pertinents sans avoir à faire une'analyse syntaxique. Cette méthode a été appliquée à différents corpus parfois non destinés au Web. Cependant, les indices linguistiques et les règles d'exploration contextuelle utilisées dans les différentes tâches ne peuvent être étendues à tout type de textes, encore moins à ceux issus du Web ([Mourad, 2001, [Coch, & Masanès, 2004]). Il est donc nécessaire de procéder à la recherche des indices linguistiques adaptés à la description des pages Web en arabe pour décrire leur contenu.

Pages Web à étudier

Une étude des pages Web des sites du corpus a permis de définir les types importants ds pages Webspour les soumettre au traitement. A partir d'un site Web, ces pages peuvent être la page d'accueil et toute page dont l'URL contient les marqueurs tels que « about », « apropos », « a_propos », « contact » …, ainsi que toute page pointée par un lien interne du même site dont le texte d'ancre contient un des marqueurs de l'ensemble suivant dans les trois langues :

about, qui sommes-nous, معلومات للاتصال، عن الموقع، للإتصال بنا، من نحن, à propos, contact, إعرف عنا، فريق العمل، إتصل بنا.(1)

Les pages Web des sites sont écrites en langage HTML, c'est donc le code HTML des pages qui sera traité. Pour chacune des pages soumises au traitement, les informations qui peuvent être étudiées à l'intériere de la page sont : le titre de la page (défini par la balise <TITLE>), les Métadonnées (les balises <META> à l'intérieur des balises <HEAD>) et le contenu textuel de la page (présent à l'intérieur de la balise <BODY>). L'outil sélectionne les parties spécifiques de la page, enlève les balises HTML, puis procède à la décomposition du texte du résultat en unités textuelles, telles que des paragraphes et les phrases de ces paragraphes. Pour simplifier, une phrase est une séquence de mots et es ponctuations (virgule et deux points) qui est suivie par un point puis un espace. Les segments textuels obtenus constituent l'espace de recherche lors du traitement qui vise à faire la classification es ces segments dans les constituants de la fiche d'archivage.

(1) Traduction respective des termes arabes en français : Informations pour contacter. A propos du site. Pour nous contacter. Qui sommes-nous. Nous connaître. Équipe de travail. Contacter nous

Filtrage et classification des segments

Dans les sites qui publient des résultats scientifiques, pour aider à compléter les informations de la fiche d'archivage, une recherche de certaines lexies et expressions peut être faites telles que :

فرضية، تجربة، وصف الطريقة، استنتاج، إشكالية، ببليوغرافيا، مقدمة، جدول المحتويات، ملخص الموضوع، خلاصة،النتائج، عرض الطريقة، تطبيق، ملخص، الغرض، الطريقة، يمكن شرحها كما يلي، تتالف من، تطبق على الشكل التالي، نستطيع ان نستنتج، تنقسم الى، هي على الشكل التالي، تتلخص بما يلي

Ou bien en français respectivement : Hypothèse, Expérience, Description de la démarche, Conclusion, Problématique, Bibliographie, avant-propos, Introduction, Table des matières, Résumé du sujet, résumé, Les résultats, Présentation de la démarche, perspective, Application, Résumé, Le but, La méthodologie, Peut-être expliquée de la façon suivante, Se compose de, S'applique de la façon suivante, Nous pouvons constater que, Se divise en, Elle prend la forme suivante, Se résume de la façon suivante.

Pour les sites Web du corpus, une étude approfondie du contenu des sites a permis de ressortir les marqueurs linguistiques et les règles pour la classification des segments textuels. Dans ce qui sent, les trois points … qui séparent deux marqueurs indiquent zéro ou plusieurs mots.

Voici donc une partie de l'ensemble des marqueurs qui caractérisent chaque classe de la fiche d'archivage. Pour un segment textuel donné, si l'un des marqueurs de la classe, qu'il soit continu ou discontinu (séparé par des …), correspond à une suite contiguë (cas d'un marqueur continu) ou non contiguë (cas d'un marqueur discontinu) des mots du segment (l'espace et la virgule sont considérées des séparateurs es mots), ce segment est alors classé dans la classe en question:

Classe «Description»: الموقع ... يصف (le site … décrit), الموقع ... يلقي الضوء , (le site … met en évidence), الموقع ... يتناول (le site … aborde) , الموقع … DESC (voir l'Annexe II).

Classe « Présentation » : الموقع ... هو (le site … est), موقع ... عبارة عن(le site … se considère), الموقع ... يتحدث عن (le site … parle ou aborde), الموقع يعرض (le site présente), الموقع يمثل (le site représente) , الموقع … PRES (voir l'annexe II).

Classe « Objectifs » : الموقع يرمي , الهدف من الموقع (le site a pour objectif) , الموقع … OBJEC (voir l'Annexe II).

Classe « Explication » : يشرح ... الموقع (le site explique),... الموقع, يفصل ... الموقع يتركب من .

Classe « Offre du site » : خدمات الموقع (les services du site), الموقع يقدم (le site offre), الموقع … OFFRE (voir l'Annexe II).

Classe « Personnes » : المسؤول عن الموقع (le responsable du site), القيم على الموقع(le tuteur du site) , صاحب الموقع(le possesseur du site) , نحن مجموعة(nous sommes un groupe), من نحن(qui sommes-nous).

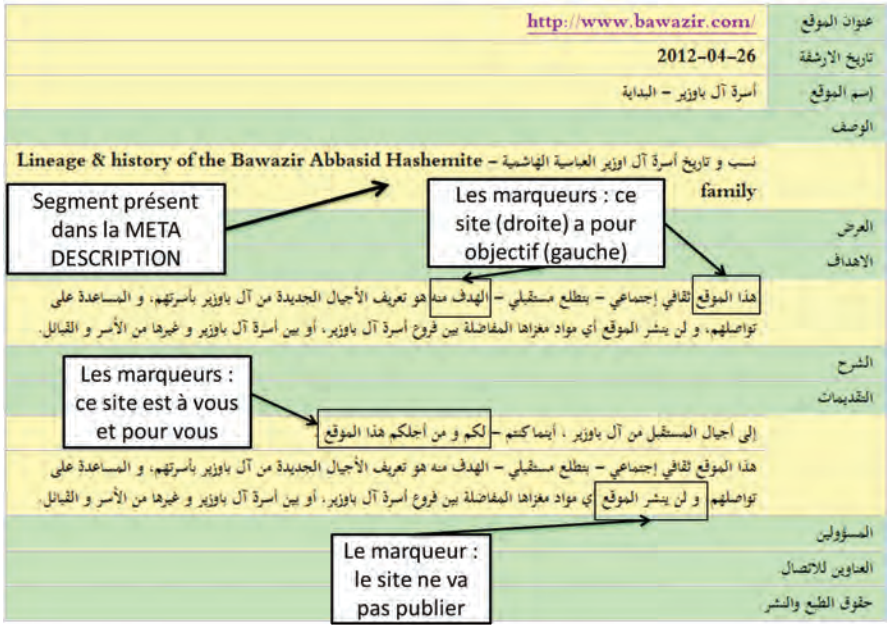
Classe « Mails » : للاتصال بنا (pour nous contacter), اتصل بنا(contacter nous).

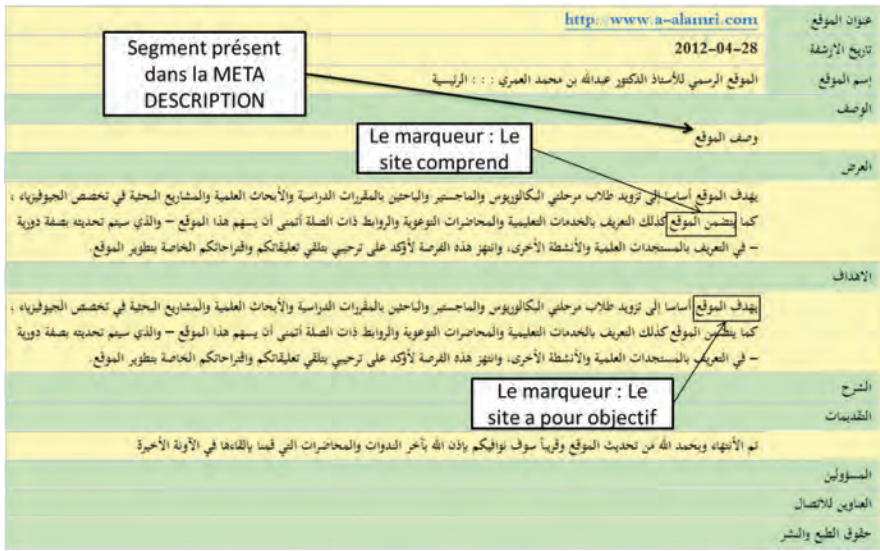
Classe « Copyright » : © , &copy, , Copyright, حقوق النشر (droit d'auteur), جميع الحقوق (tous droits réservés).

Les marqueurs définis pour les classes comptent plus de 120 marqueurs (voir pour une liste exhaustive des marqueurs pour chaque classe). Ces marqueurs sont utilisés dans l'outil avec toutes leurs variations flexionnelles, et les conjonctions « و » (et) et « ف » (car) sont optionnelles au début des marqueurs et ce pour permettre la reconnaissance des marqueurs lors d'une utilisation d'une de ces conjonctions avant le marqueur (par exemple ويعرض الموقع et فيعرض الموقع), parce qu'en arabe ces conjonctions sont collées au terme qui suit.

La sortie de l'outil est un fichier XML, quelquefois, certains constituants de la fiche peuvent reststt vide. vVoici le code XML d'une fiche d'archivage obtenue pour le site http://www.egypt.gov.eg/

Une feuille de style XSLT est utilisée pour mettre les résultats sous un format lisible. Voici des copies écrans de certains résultats obtenus :





Discussion

Les champs « Titre » et « Description » ont été renseignés pour la plupart des sites et correctement renseignés pour une bonne part, cCela s'explique par le fait que les auteurs des sites veillent à compléter le titre (qui se trouve dans la balise <title>), et à remplir la balise META DESCRIPTION dans

la balise HEAD par des informations pertinentes et ce pour vendre leurs sites auprès des moteurs de recherche. Par ailleurs le champ « copyright » a été rempli et correctement rempli pour certains sites. En revanche, les champs « présentation », «objectif», «explication», «offre du site» et « personnes, » n'ont pas été remplutpour une grande partie des sites traités. Cela est dû au fait que les auteurs des sites prennent rarement soin de ce type d'informaties. Un inconvénient de l'approche est principalement lié au problème d'ambiguïté sémantique des marqueurs dans les textes, qui demeure une des difficultés majeures et supplémentaires dans le cas des sites web. En outre, la linguistique reste, pour le moment, destinée à étudier d'une part des textes linéaires et bien structurés, et d'autre part des textes standardisés au niveau grammatical où l'étude de la phrase reste à la base de tout traitement automatique des langues, problème souligné par [Victorri & Fuchs, 1996] et d'autres, qui ont tentés de montrer la place de la phrase pour le TALN. De plus, la liberté que le langage HTML confère aux auteurs des pagsuerend les composants textuels de la pagt difficiles à classer dans l'objectif de constituer la structure logique des informations textuels de la pa, ; ce qui fait que, quelquefois, les informations d'archivage se retrouvent dispersées dans plusieurs segments textuels et ne sont donc pas classées.

Ce type d'analyse des sites Web peut être utilisé pour l'évaluation des sits, eEn effet, la plupart des analyses des sites se basent sur un schéma descriptif de certains objets comme ceux décrits dans l'article.

Bibliographie

M. Al-Hajj, M., Mourad, G.,"Extraction of reported speeches from Arabic Lebanese newspapers" Fifth International Conference on Digital Information and Communication Technology and its Applications (DICTAP) - 2015.

AL-Smadi, M., Qawasmeh O., "Knowledge-based Approach for Event Extraction from Arabic Tweets", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 6, 2016.

Alruily, M, Alghamdi, M, "Extracting information of future events from Arabic newspapers": an overview " Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)

Baccour, L., Hadrich, B., Mourad, G. (2003), «Segmentation de textes arabes en phrases basée sur les signes de ponctuation et les mots connecteurs», GEI'2003, Troisièmes Journées Scientifiques des Jeunes Chercheurs en Génie

Electronique et Informatique, Tunisie, 18-20 mars.

Coch, J., & Masanès, J. (2004), “Language engineering techniques for web archiving”. Paper presented at the IWAW-2004.

Desclés J.-P. (1997), « Systèmes d’exploration contextuelle », Co-texte et calcul du sens, (Claude Guimier), Presses de l’universitaires de Caen, 215-232.

Elsayed, H., Elghazaly, T., «Information Extraction from Arabic News”, IJCSI International Journal of Computer Science Issues, Volume 12, Issue 1, No 2, January 2015

[Feriel & Khireddine, 2015] Feriel, A., Khireddine, K., “Automatic extraction of spatio-temporal information from arabic text documents”, International Journal of Computer Science & Information Technology (IJCSIT) Vol 7, No 5, October 2015

Haettiger, M. (2003), « L’archivage des sites web d’intérêt régional » Mémoire d’étude diplôme de conservateur des bibliothèques : Bibliothéconomie : Villeurbanne, ENSSIB : 2003.

ILLIEN, G. & OURY, C. (2009), «Quelle politique documentaire pour l’archivage des sites Internet». Dans : Les collections électroniques. Une politique documentaire en mouvement, dir. P. Carbone et F. Cavalier. Paris : Éditions du Cercle de la librairie, 2009, p. 157-178.

Lounamaa, K., and Salonharju, I. (1999), «EVA - The Acquisition and Archiving of Electronic Network Publications in Finland.” Tietolinja News. Volume 1 (1999). Online. Available at: http://hul.helsinki.fi/tietolinja/0199/evaart.html. 8 October 1999.

MASANES, J. (2001), “The BnF’sproject for Web archiving”. Contribution for the European Conference on digital libraries (ECDL) 2001: What’s next for digital deposit libraries? Darmstadt, 8 Septembre 2001. [En ligne] http://bibnum.bnf.fr/ecdl/2001/france/sld001.htm

Mourad, G. (2001), «Analyse informatique de signes typographiques pour la segmentation de textes et l’extraction automatique des citations. Réalisation des applications informatiques : SegATex et CitaRE», Thèse de doctorat, Université Paris-Sorbonne, 2001.

Mourad, G. (2002), «Problématiques d’archivage et de recherche

d’informations sur le Web». Rapport du projet Watson Tunisie.

Shaalan, K., «A Survey of Arabic Named Entity Recognition and Classification.” MIT Press Journals, Cambridge, MA, USA. June 2014

VICTORRI, B., FUCHS., C., (1996). «La polysémie, construction dynamique du sens», Paris, Hermès.

Annexe I

Un échantillon du corpus, composé de 275 sites Web en aras :

www.adm.gov.ae	www.aawsat.com	www.almosmem.com	www.stars-box.com	portal.www.gov.qa
www.adpolice.gov.ae	www.abunawaf.com	www.almostshar.com	www.start10.com	www.goic.org.qa
www.deg.gov.ae	www.abyat.com	www.almothaqaf.com	www.sudaneseonline.com	www.al-nadi.com.sa
www.dgw.gov.ae	www.ahlyegypt.com	www.almountakhab.com	www.swishe.com	www.alnadwah.com.sa
www.dm.gov.ae	www.aitnews.com	www.almshaheer.com	www.syria-news.com	www.alwatan.com.sa
www.dubaided.gov.ae	www.akalaty.com	www.almustaqbal.com	www.syrian-soccer.com	www.arreyadi.com.sa
www.dubaipolice.gov.ae	www.akhbar-alkhaleej.com	www.alnadawi.com	www.taaam.com	www.okaz.com.sa
www.gia.gov.ae	www.akhbarboom.com	www.alsahaa.com	www.tabuk-news.com	www.saudidistribution.com.sa
www.moe.gov.ae	www.aklaat.com	www.alsaudeh.com	www.taibanews.com	www.se.com.sa
www.rak-traffic.gov.ae	www.aklob.com	www.alsdaqa.com	www.taleea.com	www.stc.com.sa
www.lakii.com	www.aknews.com	www.al-seyassah.com	www.tareebnews.com	www.alriyadh.gov.sa
www.iraq-amsi.com	www.al3nabi.com	www.alsh3r.com	www.alifta.net	www.amana-md.gov.sa
www.kuna.net.kw	www.alaahd.com	www.alshaab.com	www.aljamaheir.net	www.amanataljouf.gov.sa
www.annahar.com.lb	www.alafari.com	www.alshahedkw.com	www.aljazeera.net	www.armp.gov.sa
www.vdl.com.lb	www.alaflaaj.com	www.alshalan.com	www.aljazeerasport.net	www.asyah.gov.sa
www.schoolnet.edu.lb	www.alahwazvoice.com	www.alshalawa.com	www.aljoaf.net	www.baflaj.gov.sa
www.agriculture.gov.lb	www.alainteam.com	www.alshararat.com	www.alkharjonline.net	www.baish.gov.sa
www.bccl.gov.lb	www.al-akhbar.com	www.al-sharq.com	www.almagharibia.net	www.commerce.gov.sa
www.customs.gov.lb	www.alamalyawm.com	www.alsoufia.com	www.almanea.net	www.dammam.gov.sa
www.ebml.gov.lb	www.alamdi.com	www.arabiancreativity.com	www.almarefh.net	www.dmmr.gov.sa
www.economy.gov.lb	www.alankabout.com	www.arabicgospel.com	www.almeethaq.net	www.gip.gov.sa
www.finance.gov.lb	www.alanwar.com	www.arabiclinux.com	www.almersad.net	www.gosi.gov.sa
www.foreign.gov.lb	www.alapn.com	www.arabicmagazine.com	www.almohaisen.net	gate.gph.gov.sa
www.higher-edu.gov.lb	www.alarabimag.com	www.arabicnewsarchive.com	www.al-moharer.net	wmn.gph.gov.sa
www.interior.gov.lb	www.alarabnews.com	www.arabo.com	www.almostkbl.net	www.moda.gov.sa
www.isc.gov.lb	www.alasmaa.com	www.arabscoach.com	www.ahewar.org	www.moe.gov.sa
www.isf.gov.lb	www.alassil.com	vb.arabseyes.com	www.alahaideb.org	www.mof.gov.sa
www.justice.gov.lb	www.alassr.com	www.arabswe.com	www.al-aqsa.org	www.mofa.gov.sa
www.mehe.gov.lb	www.alawazm.com	www.araob.com	www.alawan.org	www.moh.gov.sa
www.ministryinfo.gov.lb	www.alayam.com	www.ar-encyclopedia.com	www.albabtainprize.org	www.mohe.gov.sa
www.ministryofdisplaced.gov.lb	www.al-ayyam.com	www.argaam.com	www.albadil.org	www.moi.gov.sa
www.moe.gov.lb	www.alazraq.com	www.arpoet.com	www.aleslah.org	www.moj.gov.sa
www.moew.gov.lb	www.albaghdadia.com	www.arraee.com	www.alghazaly.org	www.mol.gov.sa
www.moim.gov.lb	www.albawaba.com	www.e-happyfamily.com	www.alkaabi.org	www.momra.gov.sa
www.mpt.gov.lb	www.albeet.com	www.elaana.com	ar.alkarama.org	www.mopm.gov.sa
www.nna-leb.gov.lb	www.albiladdaily.com	www.elaph.com	www.almadina.org	www.mosa.gov.sa
www.omspa.gov.lb	www.al-buainain.com	www.elazayem.com	www.almajidcenter.org	www.mot.gov.sa
www.pcm.gov.lb	www.alchourouk.com	www.elhawy.com	www.almogheerah.org	www.mow.gov.sa
www.presidency.gov.lb	www.aldeerah-news.com	www.elkhabar.com	www.altaif.org	www.nbha.gov.sa
www.presidencyinfo.gov.lb	www.aldhfeer.com	www.elmuhajer.com	www.alwafd.org	www.abudhabi.ae
www.public-health.gov.lb	www.aldory.com	www.eltwhed.com	olamaa-yemen.net	www.ajman.ae
www.socialaffairs.gov.lb	www.alhawyah.com	www.elwaha-dz.com	qursansound.net	www.akhbaralarab.ae
www.southernlebanon.gov.lb	www.alhayat.com	www.el-wasat.com	rasoulallah.net	www.albayan.ae
www.state-security.gov.lb	www.alhodaif.com	www.ibnothaimeen.com	soutalhaq.net	www.alittihad.ae
www.transportation.gov.lb	www.alhololfamily.com	www.ikhwanonline.com	arabicradio.org	www.alkhaleej.ae
arabic.bayynat.org.lb	www.alhoweitat.com	www.ikhwanwiki.com	awu-dam.org	www.dubai.ae
www.almanar.com.lb	www.alhwajr.com	website.informer.com	egyleftparty.org	www.forexpros.ae
www.agar.com.ly	www.al-ilmiyah.com	www.kff.com	freearabvoice.org	www.fotosearch.ae
www.arifonet.org.ma	www.alintiqad.com	www.kfj3.com	www.almokhtsar.com	www.fujairah.ae
www.fondation.org.ma	www.al-islam.com	www.kharjhome.com	www.alitthad.com	www.pp.gov.qa
www.zayane.voila.net	www.alithnainya.com	www.khayma.com	www.aliwaa.com	www.lahyan.com
www.arabic.rnw.nl	www.manar.com	www.kooora.com	www.aljaredah.com	www.matni.com
www.manpower.gov.om	www.manutd.com	www.korabia.com	www.al-jazirah.com	www.moi.gov.qa
www.rop.gov.om	www.marxy.com	www.krokeh.com	www.alarab.com.qa	www.ktb-20.com
www.magazine.hulf.org	www.mashro3na.com	www.kslib.com	www.diwan.gov.qa	www.mohe.gov.ps

Annexe II

Les marqueurs pour la caractérisation des classes :

DESC	*PRES*	*OBJEC*	*OFFRE*
يعزز	يعُنى	يهدف إلى	يعطي
يقرر	يحتوي	الهدف منه	ينشر
يسلط الضوء	يهتم	الغرض منه	يقدم
يكرس	يعرض	يرمي	يؤمّن
يدرس	يستعرض	يسعى	يوفر
يروي	يتضمن		يخصص
يعلم			يُمكّن
يخبر			يوافي
يعرف			يساعد
ينقل			يدعو
يقص			يتطلع ليكون
يخبر			يضع بين يديك
يضم			
ينشر			
ينمي			