Ramzi Abbes[1], Émilie Guimier De Neef[2], Gilles Prigent[3]
[1]ramzi.abbes@ rd.francetelecom.com
[2]emilie.guimierdeneef@
[3]gilles.prigent@rd.francetelecom.com
France Télécom Recherche et Développement- France

TiLT : un analyseur syntaxique de surface à base de règles pour les textes arabes

Dans cet article nous allons présenter une évaluation de l'analyse de textes arabes faite avec le logiciel TiLT[1], outil de traitement automatique des textes écrits développé chez France Télécom Recherche et développement. Dans un premier temps nous exposerons les spécificités du TAL de l'arabe : le mot graphique, la voyellation et l'ambiguïté. Dans un second temps nous présenterons l'analyseur TiLT : ses ressources linguistiques et ses modules d'analyses morphologique et syntaxique de surface. Dans la dernière partie nous donnerons une évaluation du système faite sur la base de ressources linguistiques référencées.

I .Spécificités du TAL de l'Arabe

Le mot graphique en arabe est un objet complexe appelé *mot maximal,* il est décomposable en : proclitique(s), préfixe, base, suffixe(s), enclitique(s). La *base*, s'analyse en une *racine* et un *schème*. On appelle *mot minimal* l'ensemble préfixe, base et suffixe(s) [COHEN D., 1970]

La combinatoire des clitiques rend la taille du lexique potentiellement infinie.

A l'écrit, les signes diacritiques sont souvent omis. Ces voyelles sont déterminantes quant au sens des mots et à leur fonction syntaxique. En TAL, cette particularité augmente considérablement l'ambiguïté.

II . Analyseur TiLT pour la langue arabe

L'architecture de l'analyseur est la suivante :

Texte -> segmentation -> analyse lexicale/analyse morphologique -> analyse syntaxique de surface.

III . Données lexicales

La construction du lexique part des formes les plus "primitives", racine et schème, et décline tous les mots minimaux voyellés possibles en arabe en suivant des modèles de conjugaison pour les verbes et des modèles de déclinaisons pour les noms. Les autres éléments (mots outils, proclitiques…) sont introduits directement avec leurs formes finies.

Lors de ce processus tous les phénomènes graphiques et phonétiques dus à la fusion de la racine et du schème ou encore à la concaténation des infixes à la base sont traités. Nous obtenons donc un lexique de mots minimaux complètement voyellés.

IV . Analyse lexicale

Pour tenir compte de l'omission fréquente des signes diacritiques à l'écrit l'analyse lexicale peut se faire en mode tolérant à l'absence de voyelles. Nous pouvons aussi prendre en compte les pratiques d'écriture comme la confusion de la "Hamza" et du "Alif" au début des mots ou encore le "ya" et la "Alif maksoura" à la fin du mot [ABBES R., DICHY J. and HASSOUN M., 2004b]

V . Analyseur morphologique

Les mots qui ne sont pas reconnus par l'analyse lexicale sont traités par le composant morphologique, essentiellement des mots maximaux contenant des proclitiques. Le système ne retourne que les découpages autorisés en langue arabe, par un ensemble de règles qui vérifient la compatibilité entre le proclitique supposé et le mot restant.

VI . Analyse syntaxique de surface

A l'issue des phases précédentes, nous nous trouvons souvent avec une multitude de solutions pour chaque mot. Hors contexte et sans voyellation, il est rare d'obtenir une solution optimale, même pour un lecteur humain. Une exploration des contextes les plus proches peut lever certaines ambiguïtés. Pour cela nous utilisons des grammaires exploratoires de flux du type *chunk*.

---

[1] Le terme TiLT n'est pas une marque déposée mais un nom de code interne du logiciel

Le *Chunking* consiste en la segmentation des textes en syntagmes nominaux ou verbaux non récursifs. Lors d'un chunking les mots sont traités dans leur ordre d'arrivée et les groupes sont sélectionnés en fonction de leurs tailles, en favorisant les groupes les plus longs. Nous fixons des contraintes d'accord entre les éléments d'un même groupe et des contraintes de succession entre les groupes.

Les grammaires légères permettent de lever une partie des ambiguïtés entre les catégories grammaticales, par exemple entre un verbe et un nom concurrents sur une même forme graphique. Malgré l'absence de voyellation en arabe nous obtenons souvent des solutions correctes, mais nous restons confrontés au choix de la bonne solution. Le traitement local atteint ses limites quand l'accord doit se faire entre des mots de groupes différents. La levée totale d'ambiguïté nécessite souvent des considérations sémantiques, ce que nous n'avons pas encore intégré au traitement de l'arabe.

VII . Évaluation

Pour conclure cet article nous donnerons une évaluation des analyses de TiLT. Pour cela nous nous baserons sur un corpus de dépêches journalistiques étiquetées. Le corpus comporte 1 000 000 de mots, la validation des analyses morpho-syntaxiques (levée d'ambiguïtés) a été faite manuellement.

L'évaluation se fera en deux étapes. La première s'intéressera à l'analyse du texte voyellé pour examiner l'efficacité de la grammaire. La seconde s'intéressera à l'analyse du même texte mais sans les voyelles pour examiner les capacités de revoyellation et de levée d'ambiguïté du système.

VIII . Bibliographie

[COHEN D., 1970] COHEN DAVID. Essai d'une analyse automatique de l'arabe. In: David Cohen. Etudes de linguistique sémitique et arabe. Paris:Mouton, 1970, pp. 49-78.

[ABBES R., DICHY J. and HASSOUN M., 2004a] ABBES RAMZI, DICHY JOSEPH and HASSOUN MOHAMED. The Architecture of a Standard Arabic Lexical database: some figures, ratios and categories from the DIINAR.1 source program. COLING'04. Proceedings of the Workshop Computational Approches to Arabic Script-bases Languages., 28 august 2004., Genova. pp 15-22.

[ABBES R., DICHY J. and HASSOUN M., 2004b] ABBES RAMZI, DICHY JOSEPH and HASSOUN MOHAMED. Les mots arabes dans un corpus journalistique contemporain: statistique lexicales et pratiques d'écriture. In: Brahim Ben Mraad. colloque international de lexicographie : "dictionnairique et corpus, 19-21 juin 2004, Tunis: Tunisie.

[BUCKWALTER T., 2004] BUCKWALTER TIM. Issues in Arabic Orthography and Morphology Analysis. COLING'04, Proceedings of the Workshop Computational Approaches to Arabic Script-bases Languages. 28 aout 2004., Genova. pp 31-41.

[DICHY J., 1990] DICHY JOSEPH. L'Ecriture dans la représentation de la langue : la lettre et le mot en arabe. Thèse pour le doctorat d'état (ès Lettres). lyon:Université Lumière-Lyon 2, 1990,

[GAUBERT C., 2001] GAUBERT CHRISTIAN. Stratégie et règles minimales pour un traitement automatique de l'arabe. Thèse en études Arabes. Marseille:Aix-Marseille I - Université de Provence, 2001, 433 p.

[GUIMIER E, 2002] GUIMIER EMILIE. L'analyse de textes : entre données linguistiques et robustesse. Journée METIL, Métiers des Industrie de la Langue, 7 mars 2002, Paris, http://www.apil.asso.fr/metil.htm.

[GUIMIER E, BOUALEM M, CHARDENON C et al, 2002] EMILIE GUIMIER DE NEEF, MALEK BOUALEM, CHRISTINE CHARDENON, PASCAL FILOCHE, JÉRÔME VINESSE, Natural Language Processing software tools and linguistic data developed by France Telecom R&D, IEMCT : Indo European Conference on Multilingual Technologies, Pune, Inde, 2002