*Une approche de désambiguïsation morpho_lexicale évaluée sur l'analyseur morphologique Alkhalil**

K.Z Bousmaha[1,2], S. Charef_Abdoun[1], L. Hadrich_Belguith[2], M.K Rahmouni[1]

[1]Université d'Oran, Faculté des sciences, Laboratoire RIIR, Algérie
[2]ANLP Research Group, Laboratoire MIRACL, Université de Sfax, Tunisie

kzbousmaha@yahoo.fr , l.belguith@fsegs.rnu.tn, mk_rahmouni@yahoo.fr

Abstract: *Out of any context, most words have several meanings and several functions. The lexical disambiguation, for instance, consists of selecting the right meaning of a polysemic word in a given context. Several methods and approaches exist at all level of disambiguation : could it be morphological, lexical or semantic (in case of language processing) but as far as Arabic is concerned, ambiguity persists because of the non-diacritisation of words. In order to considerably reduce the ambiguity rate, we propose in this article a disambiguation approach based on the selection of the right diacritics at different analyses stages. This approach combines between a linguistic approach and a multicriteria decision one and could be regarded as a choice alternative to solve the morpho-lexical ambiguity problem regardless of the diacritics rate of the processed text. As to its evaluation, we have tried the disambiguation of the on-line Alkhalil morphological analyzer (the proposed approach can be experienced on any morphological analyzer of arabic language). Alkhalil Morpho Sys, 1.3, 2011 is an open source. We have obtained encouraging results with an F-measure of more than 80 percent.*

Résumé : *Hors contexte, la plupart des mots ont plusieurs sens et plusieurs fonctions. La désambiguïsation lexicale, par exemple, consiste à choisir la bonne signification d'un mot polysémique dans un contexte donné. Plusieurs méthodes et approches existent à tous les niveaux de désambigüisation: morphologique [7], lexicale [3], sémantique [4] pour le TAL (traitement automatique des langues). Mais pour la langue arabe l'ambigüité s'accentue par la non diacritisation des mots. Afin de réduire considérablement ce taux d'ambigüité, nous proposons dans cet article une approche de désambigüisation qui se fonde sur le choix des bonnes diacritiques lors des différentes analyses. Cette approche combine une approche linguistique à une approche multicritère d'aide à la décision. Cette combinaison peut être considérée comme une alternative de choix pour remédier au problème de l'ambigüité morpho_lexicale quelque soit le taux de diacritiques du texte traité. Pour son évaluation, nous avons essayé la désambigüisation de l'analyseur morphologique on-line Alkhalil (L'approche que nous proposons peut être expérimentée sur n'importe quel autre analyseur morphologique da la langue arabe). Alkhalil Morpho Sys, Version 1.3, 2011, un open source http://www.alecso.org.tn/index.php?option=com_content&task=view&id=1302&Itemid=956&lang=a . Nous avons obtenu des résultats encourageants avec un F-Measure de plus de 80%.*

Keywords: *TALA, Alkhalil morphological analyzer, disambiguation method, diacritisation, Approach to Multicriteria Decision (AMD), segmentation, contextual exploration, tagging, augmented transition networks (ATN).*
Mots clés *: TALA, analyseur morphologique ALKHALIL, méthode de désambigüisation, diacritisation, méthode multicritère d'aide à la décision (AMD), segmentation, exploration contextuelle, étiquetage, Réseaux de transition augmenté(ATN).*

**A morpho-lexical disambiguation approach using Alkhalil morphological analyser*

© Revue RIST, vol. 20, n° 2, p.32-46

1 Introduction

Plusieurs analyses grammaticales, c'est l'obstacle à la compréhension des textes. La nécessité d'une méthode de désambiguïsation permet de créer un système performant, robuste, rapide et moins ambigu pour une analyse grammaticale correcte. Pour ce faire, différentes méthodes existent,

Il existe des méthodes de désambiguïsation lexicales qui reposent sur des connaissances issues de ressources lexicales/sémantiques (knowledge-based WSD), citons l'algorithme de LESK, de LESK simplifié, l'algorithme de Yarowsky (1992). Beaucoup d'inconvénients sont inhérents à ces méthodes à savoir l'indisponibilité de ressources lexico-sémantiques pour beaucoup de langue et possibilité d'explosion combinatoire si l'on essaie de désambiguïser tous les mots du texte [3][6]. Il y a celles qui reposent sur les approches supervisées, nécessitant des corpus d'entraînement étiquetés manuellement très couteuses en temps et en argent avec possibilité de goulot pour l'acquisition de données et, d'autre part, des approches non-supervisées plus intéressantes à savoir les approches non supervisées classiques (clustering) qui exploitent les données non annotées ; et d'autre part les approches à base de savoirs qui utilisent des connaissances issues de ressources lexicales [16]. Des systèmes hybrides, sont aussi apparus, combinant plusieurs sources d'information (fréquence des mots, informations lexicales, syntaxiques, contextuelles, etc.). Pour la qualité de leurs résultats et la détermination des sens des mots , ils proposent deux solutions: (i) l'enrichissement automatique des réseaux de type WordNet pour y introduire les informations permettant de répondre aux critiques formulées à leur encontre; (ii) l'extraction des sens des mots automatiquement à partir de corpus, sans utiliser de dictionnaires existants (désambiguïsation sémantique automatique) [24].

Pour l'ambigüité morphologique [7], Il existe des méthodes à base des modèles mathématiques et stochastiques (probabiliste et statistique), d'autres basées sur des modèles par contraintes, à base de règles, et d'autres basées sur la théorie décisionnelle permettant le classement multicritère des scénarios de désambiguïsation afin de déterminer le meilleur par un classement des différents critères d'évaluation, en bâtissant un système fiable et réduire le nombre d'interprétation issues de l'analyse morphologique[14]. Nous nous intéressons ici à ces dernières.

La principale caractéristique de la langue arabe vient du fait qu'il s'agit d'une langue agglutinée et surtout voyellée, où sans diacritisation, il est difficile de distinguer le sens et la fonction des mots. Cette caractéristique introduit, de fait, une forte ambiguïté avec laquelle il va falloir « jouer » comme l'a dit Debili [10] dans le cadre du traitement automatique de la langue. De ce fait, notre approche de désambiguïsation est fondée sur le choix des bonnes diacritiques lors des différentes analyses pour soulever l'ambigüité du mot et de reconnaitre ainsi ses caractéristiques grammaticales et lexicales.

Après un bref aperçu donné en section 2 sur quelques approches de désambiguïsation adaptées à la langue arabe, nous présentons dans la section 3 de cet article notre approche ainsi que l'analyseur auquel nous avons intégré notre proposition, Nous détaillons dans la section 4 les différents modules de notre analyseur D_Alkhalil en illustrant notre approche par un exemple. Nous décrivons dans la section 5, les expérimentations réalisées pour évaluer notre travail et les résultats obtenus de l'analyseur morphologique avant et après l'ajout de notre solution.

2 Bref aperçu des approches de désambiguïsation de la langue arabe

Un problème essentiel lors de la désambiguïsation est la manière d'adapter pour l'arabe les méthodes de désambiguïsation existantes, on trouve l'emploi de:

- SVM (Support Vector Machines) utilisés pour la segmentation des mots et de l'étiquetage. L'étiquetage a été appliqué aux mots segmentés en utilisant une segmentation à partir du corpus annoté (Mona Diab); [12]
- MMC (Modèle de Markov Caché) et Modèle de Markov bidimensionnel pour la segmentation et l'étiquetage de l'arabe appliqués sur des corpus non étiquetés utilisant l'algorithme BaumGallois ; [20]
- Modèle de langue de n-gram décrivant un système de segmentation de mots pour l'arabe ;
- L'étiquetage à base des règles (Brill, 95 voir référence 4), le modèle entropie maximal de Ratnaparkhi 1996;
- Calcul des probabilités qui permet d'assigner une probabilité à chaque séquence d'étiquettes potentielle, et le module de désambiguïsation permettant de calculer la séquence d'étiquettes la plus probable pour chacune des séquences du texte à analyser. [25]

Nous reprochons à ces approches l'utilisation d'une grande masse de données annotées (pour un étiquetage supervisé) ou un lexique qui inscrit toutes les étiquettes possibles pour chaque mot (pour un étiquetage non supervisé). Même si la tâche est loin d'être résolue, de nombreux travaux portent sur ce domaine et les pistes explorées sont multiples.

L'une des pistes à explorer et que nous avons retenue, est la désambiguïsation par la prise en compte des diacritiques lors des différentes analyses quelque soit le taux de diacritiques du texte traité. Trouver les bonnes diacritiques d'un mot peut être une bonne voie pour le désambiguïser sachant que le sens et la fonction grammaticale d'un mot sont fortement liés à sa bonne diacritisation.

3 Principe de notre approche

Notre approche de désambigüisation se fonde sur le choix de la bonne diacritisation du mot du texte du fait que le mot arabe en accepte plusieurs. Bien diacritiser revient à choisir en contexte la bonne diacrité d'un mot afin d'en déterminer le sens et la fonction. La problématique est double (i) Comment restituer les diacritiques potentielles de chacun des mots d'un texte analysé morphologiquement, alors que plusieurs catégories grammaticales peuvent être affectées à un mot et (ii) Comment choisir le bon schème diacritique parmi tout un ensemble proposé pour une même catégorie grammaticale attribuée à ce mot? La tâche de désambiguïsation semble alors difficile.

Notre choix s'est penché en premier lieu sur une méthode multicritère d'aide à la décision à base de TOPSIS [15] (Technique for Order by Similarity to Ideal Solution) du fait de sa robustesse et de son fondement mathématique. La méthode TOPSIS est développée par Hwang et Yoon en 1981, son fondement consiste à choisir une solution qui se rapproche le plus de la solution idéale, en se basant sur la relation de dominance qui résulte de la distance par rapport à la solution idéale (la meilleure sur tous les critères) et de s'éloigner le plus possible de la pire solution (qui dégrade tous les critères). Il s'agit de

réduire le nombre de scénarios[1] de désambiguïsation, et de classer les scénarios efficaces selon leurs scores globaux calculés.

L'intérêt d'adopter une approche multicritères pour lever l'ambiguïté morphologique dans le T.A.L.A., peut être résumé en deux (02) points [21]:

- La réduction de l'ensemble des étiquettes (scénarios) candidates: elle permet de réduire d'emblée le nombre d'étiquettes de correction, en éliminant ceux dominées, générant ainsi l'ensemble des étiquettes efficaces ;
- La classification des étiquettes efficaces, selon un score global obtenu après traitement suivant un ordre décroissant.

Mais les scénarios associés à un mot donné isolément ne peuvent être désambigüisés et l'ensemble des scénarios probables de ce mot ne peut être réduit quelques soient les critères de décision choisis du fait que:

(1) Le mot à traiter ne doit pas être pris séparément de son ensemble. Nous devons prendre en compte la fenêtre contextuelle des unités lexicales reliées syntaxiquement (encore plus autour du mot ambigu) qui doit être de taille variable. Ceci nous aidera à détecter le type de l'ambiguïté qui caractérise le mot et enfin déduire sa catégorie et/ou son trait grammatical (e).

Exemple 1: **كتب كثيرة في المكتبة**

La non diacritisation génère plusieurs cas d'ambiguïtés lexicales et morphologiques. Le mot non diacrité **كتب** [ktb] possède 16 diacritisations potentielles, représentant 9 catégories grammaticales différentes. [10]:

« / kataba » (Il a écrit) ; « / kutiba » (Il a été écrit) ; « / kutub » (des livres) ; « / katb » (un écrit) ; « / kattaba » (Il a fait écrire) ; « / kuttiba » (faire écrire - forme factitive) ; « / kattib » (fais écrire) etc ..

(2) Pour une même catégorie et/ou trait grammatical(e) d'un mot donnée dans une phrase donnée, plusieurs diacritisations sont possibles et donc plusieurs sens.

Exemple 2:

Unité lexicale	1ere interprétation	2 éme interprétation	3 éme interprétation
مدرسة	مَدرَسَة École	مُدَرَسَة Enseignée	مُدَرِسَة Enseignante

On voit bien à travers ces exemples qu'un mot ne peut être pris séparément de son contexte et qu'un texte ne peut être assimilé à un sac de mots mais plutôt à un ensemble fortement structuré de termes qui permettent de communiquer des informations d'une grande précision. D'où la nécessité d'une approche linguistique [5] [8] [13] précédant la méthode multicritère choisie.

Notre approche se compose de deux étapes:

1. Nous commençons par une analyse linguistique**.** Après segmentation en phrases**,** une analyse morpho_syntaxique est effectuée dont l'objectif est d'associer à chaque unité lexicale sa catégorie grammaticale (nom, verbe, adjectif...). Le principal intérêt de cet étiquetage est qu'il permet d'opérer un premier traitement de désambiguïsation des mots. L'étiqueteur utilisé peut ainsi associer plusieurs étiquettes (catégories) à chaque

[1] Désormais, dans notre approche, les scénarios seront assimilés aux schèmes

unité lexicale (surtout pour un mot ambigu), la deuxième étape consiste en une application de règles de grammaire implémentées sous forme d'un ensemble de réseaux de transitions augmentés (ATN) sur le texte préalablement étiqueté. Il doit par conséquent choisir parmi toutes les catégories affectées précédemment à une unité lexicale celle qui correspond au mot dans la phrase considérée, en s'appuyant sur l'ATN correspondant. Une fois le trait grammatical associé, une deuxième désambigüisation est ainsi faite, les (ou quelques) diacritiques les plus évidentes de ce mot sont trouvées et l'ensemble des schèmes candidats associé au mot est ainsi réduit. En vue d'améliorer les performances du système en termes de qualité de solutions et de temps d'exécution, et afin de prendre en considération ces problèmes d'ambiguïtés au cours de l'analyse de la langue, il n'est plus à démontrer qu'une architecture multi-agent de ses différents modules est à envisager (beaucoup d'analyseurs pour le TALA adoptent cette stratégie : mouhalil[2], maspar[3] etc..)

2. Ensuite, nous utilisons les performances formelles d'une AMD pour filtrer parmi ces schèmes candidats lesquels sont effectivement retenues et déterminer ainsi et si possible, les diacritiques finales de ce mot (son sens).

Afin d'évaluer notre approche, nous nous sommes penchés sur les outils de traitement automatique de la langue arabe qui sont l'ensemble des recherches et développements visant à modéliser et reproduire, à l'aide de machines, la capacité humaine à produire et à comprendre des énoncés linguistiques dans des buts de communication.

Notre choix s'est porté sur Alkhalil Morpho Sys qui est un analyseur morphologique on line pour le texte arabe standard. Alkhalil peut traiter les textes non diacrités, ainsi, il peut traiter les textes diacrités partiellement ou totalement. Nous avons choisi Alkhalil (Alkhalil Morpho Sys) car il pourrait être considéré comme le meilleur système morphologique arabe. En fait, Alkhalil a remporté la première position, parmi les 13 systèmes morphologiques arabes à travers le monde, à un concours organisé par la Ligue arabe pour l'éducation la culture et la science (ALECSO) (برنامج الخليل الصرفي 2010).

Il est basé sur la modélisation d'un très large ensemble de règles morphologiques arabes[4], et sur l'intégration de ressources linguistiques qui sont utiles à l'analyse. Malgré les performances de cet analyseur, et de la plupart des analyseurs de la langue arabe, nous leur reprochons l'absence du module de désambigüisation, le nombre de sorties attribuées à chaque mot du texte à analyser est considérable.

De ce fait, nous avons intégré à Alkahlil, un module de désambigüisation morpho_lexicale à base de notre approche et nous avons baptisé l'analyseur ainsi obtenu D-Alkhalil.

---

[2] (Haddad et al, 2007) Haddad A., Ben Ghezala H., Ghenima M. :"Conception d'un catégoriseur morphologique fondé sur le principe d'Eric Brill dans un contexte Multi-Agents", 26th conference on Lexis and Grammar, Bonifacio, 2-6 October 2007

[3](Aloulou *et al*, 2003), Aloulou C. , Belguith_Hadrich L. , "Analyse syntaxique de l'Arabe : le système MASPAR", *RÉCITAL 2003*, Batz-sur-Mer, 11-14 Juin 2003.

[4]Alkhalil contient environ 7000 racines obtenues à partir de Sarf ,(sarf 2007, an open source Arabic morphology system http://sourceforge.net/projects/sarf/) et NEMLAR corpus[1]. NEMLAR : Network for Euro-Mediterranean LAnguage Resources. Ce corpus a été produit dans le cadre du projet NEMLAR. Le corpus écrit NEMLAR est constitué de 500 000 unités lexicales regroupés en 13 catégories différentes, visant à obtenir un corpus bien équilibré qui offre une représentation de la variété de traits syntaxiques, sémantiques et pragmatiques de la Langue arabe moderne. Chaque racine est reliée avec des procédures de dérivation spécifiques utilisées pour calculer les mots de cette racine.

4 Architecture de L'analyseur D_Alkhalil

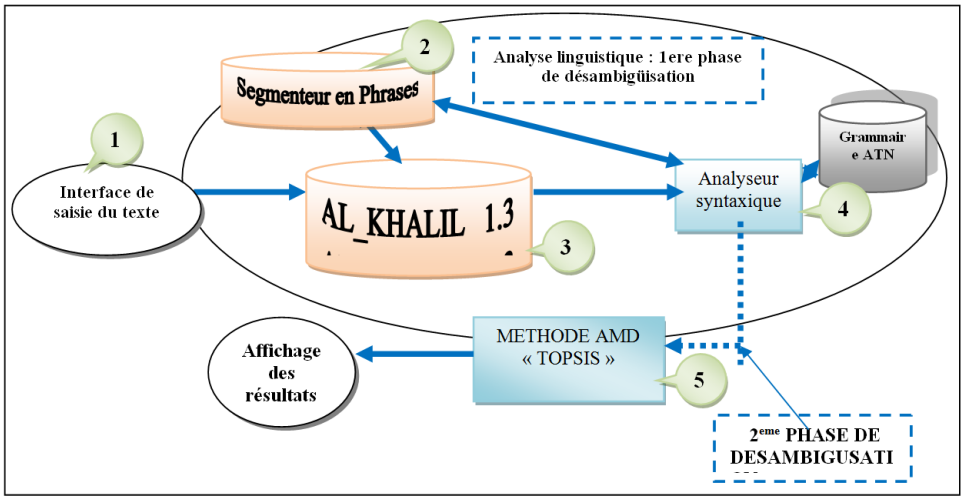

Figure 1: Architecture D_Alkhalil

(1) Entrée du texte dans Alkhalil;

(2) Alkhalil procède à une segmentation en mots et non en phrases. De ce fait, nous avons intégré l'outil $STAr_+$[5] comme un module pour la segmentation du texte en phrases, STAr (Segmenteur de Textes Arabes) et $STAr_+$ ont été réalisés au sein de notre équipe ANLP group, l'outil s'appuie sur la méthode de l'exploration contextuelle qui prend en compte le contexte droit et gauche de chaque marqueur/déclencheur (signes de ponctuation, les conjonctions de coordination et certains mots outils) passible de segmenter en phrase. L'outil compte 183 règles de segmentation en phrases et propositions. À la rencontre de la virgule par exemple, nous ne pouvons décider parfois de la fin ou pas de la proposition, nous faisons appel à Al_khalil. Des informations morphologiques sur le contexte droit et/ou gauche de la virgule sont fournies et mémorisées (pour le traitement suivant). Parfois, même l'analyseur syntaxique est sollicité lors de la segmentation d'un mot ambigu. L'évaluation du système STAr/STAr+ ont montré un taux de rappel de 94.78% et un taux de précision de 93.14%;

(3) Alkhalil fait une analyse sur les données reçues (analyse morpho-lexicale, analyse syntaxique….). L'analyse morphosyntaxique est réalisée en cinq étapes:

a. Prétraitement

Il procède à la segmentation de la phrase en mots, puis a leur normalisation en supprimant à la fois « ashida » et les diacritiques (si elles existent). Il enregistre en mémoire une copie complète de diacritiques réelles des mots d'entrées, afin de rejeter les résultats d'analyse incompatibles avec ce mot diacrité.

[5] STAr : Segmentation de textes arabes basé sur l'analyse contextuelle des signes de ponctuations et de certaines particules, L. Belguith Hadrich, L. Baccour, G. Mourad, Actes de la 12 ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'2005), Vol. 1 pp. 451–456, Dourdan-France, 6 – 10 Juin 2005. Il a été renforcé par l'ajout d'autres règles et reprogrammé en Java (Star+ : K.Hassaine, K.Z.Bousmaha, L.Belguith Hadrich, mémoire de master, Université d'Oran, Faculté des sciences, Département d'informatique 2011)

b. Segmentation

Cette étape porte sur le mot orthographique obtenu après prétraitement. Le système le considère comme une série de constituants (proclitiques + tige + enclitique). Il propose toutes les segmentations imaginables en passant par les listes proclitiques et enclitique définies. Une vérification de la compatibilité entre proclitiques et enclitiques résultante de chaque segmentation est faite.

c.Analyse de la racine (tige)

La même tige peut conduire à des interprétations diverses. Le système procède à une analyse en trois phases: Une première interprétation correspond à un mot non dérivable, une seconde interprétation peut se référer à un nom (un substantif dérivé) et une troisième à un verbe.

Pour chaque phase, la validation est effectuée en tenant compte des critères de compatibilité des proclitiques et enclitiques avec la tige et les informations morphologiques de toutes les segmentations valides sont fournies.

d. Dépistage des résultats

Les résultats obtenus à partir de l'analyse précédente seront soumis aux processus de sélection par exemples: la concordance entre proclitiques et enclitiques avec les caractéristiques de sorties syntaxiques, la concordance de l'allographe hamza (ء, أ, إ, ئ, ou ؤ) dans les solutions proposées par le système avec celui du mot en entrée, la concordance des diacritiques des solutions proposées par le système avec ceux qui peuvent exister dans le mot d'entrée.

e.Affichage des résultats de l'analyseur morphosyntaxique

Alkhalil Morpho Sys permet ainsi l'identification de l'ensemble des solutions possibles pour un mot donné. Les sorties de l'analyse de mots arabes sont présentés dans un tableau exhaustif qui montre les racines possibles et les schèmes correspondants entièrement diacrités; leurs catégories grammaticales, ses proclitiques et enclitiques.

L'analyse morphologique se doit de reconnaitre toutes les segmentations licites et associer à toutes les unités lexicales qui en sont issues leurs diverses diacritiques potentiels. La contribution de l'analyse morphologique au processus de diacritisation, consiste à l'élimination de certaines diacritiques au travers l'analyse des formes agglutinées. La résolution peut être atteinte dans certains cas, dans d'autres cas, les catégories grammaticales proposées pour un mot donnée peuvent être très nombreuses d'où la nécessité d'une phase de désambigüisation.

(4) Analyseur syntaxique, une phase de désambiguïsation: L'analyseur se base essentiellement sur la technique des réseaux de transition augmentés (ATN). Les ATN sont représentés habituellement par un graphe, ils ont la capacité de prendre des notes en cours d'opération et de se référer à ces notes pour prendre des décisions ultérieures décrivant un état et des arcs permettant de passer d'un nœud à un autre, avec l'enregistrement des états déjà empruntés dans des registres avec flags. Ceci est le point fort de ce genre de réseaux. « Les ATN n'imposent pas de restrictions formelles, mais par son intérêt pratique et calculatoire. Le formalisme des ATN peut être utilisé pour décrire des dépendances syntaxiques assez compliquées et profondes surtout pour un système récursif, de façon relativement intuitive et facile à implémenter.»[23].

Les ATN sont lisibles, compréhensibles, rapides, performants et modulaires. L'utilisation de la technologie à état finis à été le sujet d'étude dans plusieurs travaux de recherche et privilégié par différentes équipes de recherche tels Xeros et Bell labs research center pour créer des applications capables d'aborder les différents niveaux d'une analyse linguistique.

Les ATN implémentés dans notre analyseur utilisent une base de règles qui comporte deux (02) catégories de règles:

- **Une première catégorie**, qui intervient directement sur les schèmes associés aux catégories grammaticales potentielles d'un mot (classe spécialisée), elle est à base de règles contextuelles et d'heuristiques. Par un passage superficielle de la phrase, elle décide d'emblée des diacritiques les plus évidentes des schèmes potentiels du mot.

  Exemples de règles contextuelles

  Règle 1: Après une particule de subordination (ou du génitif) « حـرف جـر» vient toujours un nom génitif «اسـم مـجـرور».

  Règle 2: Après une particule du subjonctive (ou de l'accusatif) « حـرف نـصـب» pour un nom vient toujours un nom subjonctif «اسـم مـنـصـوب».

  Règle 3: Après une particule du subjonctive (ou de l'accusatif) « حـرف نـصـب» pour un verbe à l'inaccompli vient toujours un verbe à l'inaccompli « فـعـل مـضـارع».

  Ce premier passage nous permet de réduire justement et sans ambigüité le nombre de schèmes associés aux différentes catégories grammaticales affectés au mot ambigu.

- **Une deuxième catégorie de règles** concerne l'ordonnancement des mots dans la phrase et l'attribution à chacun des mots du texte la catégorie qui est la sienne dans le contexte où ce mot apparaît. Elle concerne la catégorie du mot (la classe générique). Si la phrase n'a pas été reconnue par l'un des ATN de la base, elle est rejetée.

Au travers les étapes ultérieures, nous constatons que dans plus de 72% des cas, la catégorie grammaticale est trouvée. Ce qui signifie que l'étiquetage grammatical amène à une amélioration des résultats liés à la diacritisation. Le nombre des schèmes potentiels étant réduit, la résolution étant obtenue s'il n'en subsiste qu'un seul, sinon nous passons à une autre phase de désambigüisation. Cette phase est nécessaire (dans la plupart des cas) que pour la diacritisation du mot.

(5) Une autre phase de désambiguïsation: cette phase permet de réduire le nombre d'interprétations issues des premières analyses grâce à la démarche multicritères à base de la méthode TOPSIS [15].

Topsis consiste à choisir une solution qui se rapproche le plus de la solution idéale (la meilleure sur tous les critères) et de s'éloigner le plus possible de la pire solution qui dégrade tous les critères. Cette méthode de six étapes se base sur la relation de dominance qui résulte de la distance par rapport à la solution idéale.

On définit X = {x1, x2, x3, …, xn} l'ensemble des scénarios de correction. Ces scénarios sont différents en nombre fini et constituent l'intégralité des solutions possibles.

Pour choisir le meilleur scénario de « X », on utilise un ensemble F = {f1, f2, f3, …, fn} qui constitue une famille cohérente de critères.

On passe par une succession d'étapes à commencer par dresser la liste des actions potentielles, dresser la liste des critères, définir une fonction d'évaluation (fonction de performance), établir un tableau de performance, pondération et agrégation des performances et enfin classification des scénarios.

Il est toutefois nécessaire de pondérer les critères, il existe des pondérations globales tels : Normal, Gfldf, Idf, et Entropie et comme il existe aussi des pondérations locales [21].Notre choix à été porté sur une pondération globale : l'entropie ou l'incertitude moyenne car elle est la seule méthode qui tient compte de la distribution des unités lexicales dans le texte. Cette méthode est une technique objective de pondération des critères, l'idée est qu'un critère « j » est d'autant plus important que la dispersion des évaluations des actions est importante. Ainsi les critères les plus importants sont ceux qui discriminent le plus entre les actions.

Notre choix s'est fait sur une pondération globale, car nous avons constaté qu'il fallait pondérer localement l'importance du schème dans sa catégorie grammaticale, et globalement mesurer la représentativité de l'unité lexicale dans le texte.

Le problème de la détermination des critères cohérents c'est-à-dire: exhaustifs, non redondants et formant une cohésion entre eux [18] ne s'avère pas une tâche facile. Pour discriminer entre les scénarios, nous avons défini 2 critères: La juxtaposition des diacritiques, ce critère va utiliser la position des diacritiques pour lever l'ambiguïté, il s'agit d'un critère à maximiser autant que le deuxième critère où il s'agit de calculer la fréquence d'apparition de chaque scénario candidat dans le texte.

Pour mieux comprendre le principe de la méthode, nous allons l'illustrer par un exemple. **Exemple:**

Soit la phrase Ph d'un texte T qui se trouve à l'entrée de l'outil D_Alkhalil, après une première étape de traitement linguistique déjà explicitée, la phrase passe par une deuxième phase de désambigüisation en suivant les étapes de l'AMD:

Ph = «ذهــب الطفل إلى البستان », Analysons le mot « ذهــب »

<u>Étape 1</u>: détermination des scénarios

Une fois l'analyse linguistique faîte, on obtient un ensemble réduit de schèmes pour ce mot (le scénario probable « nom » comme catégorie grammaticale à été éliminé par les étapes antérieures). Dans ce cas l'ensemble E considéré comme l'ensemble des scénarios probables pour ce mot sera:

E= {Verbes: فَـعَـلَ, فَـعِـلَ, فَـعُـلَ, فَـعِـلّ, فُـعِـلَ}.

<u>Étape 2</u>: Application des critères sur les scénarios

- **Critère 1 :** juxtaposition des diacritiques :
  A chaque diacritique est affectée soit le chiffre 0 soit 1 selon sa bonne position ou pas dans le schème.

Scénarios :{ فَـعَـلَ 1 1 1, فَـعِـلَ 1 0 1, فَـعُـلَ 1 0 1, فَـعِـلّ 1 0 0, فُـعِـلَ 0 0 1,فَـعَـلْ 1 1 0}

- **Critère 2 :** Fréquence d'apparition. (PA,PB,….. paragraphes du texte)

  A chaque scénario probable nous comptabilisons sa fréquence d'apparition au sein de chaque paragraphe du texte (sans la diacritique de la dernière lettre car c'est une marque casuelle qui dépend du trait grammatical (la position) du mot)

Exemple:

Pour le Scénario فَـعَـلَ nous comptons 2 fois en PA, 1 fois en PB, …3 fois en Pj

<u>Étape 3,4</u> : Appliquer de la fonction d'évaluation et Générer la matrice d'évaluation

Critère	فَـعَـلَ	فَـعِـلَ	فَـعُـلَ	فَـعِـلٌ	فُـعِـلَ	فَـعَـلُ
juxtaposition des diacritiques	**1+1+1=3**	**2**	**2**	**1**	**1**	**2**
Fréquence d'apparition	**16**	**5**	**3**	**5**	**5**	**16**

Tableau 1: Matrice d'évaluation à partir du texte

Exemple: Pour le scénario 1 فَـعَـلَ, nous lui attribuons dans la matrice d'évaluation (tableau1) le nombre 3 qui n'est que 1+1+1 (voir étape 2, critère 1)

<u>Étape 5</u>: Agrégation des performances et pondération des critères

- **Normalisation de la matrice d'évaluation :** e'i,j
  Cette normalisation est faite en appliquant la formule[6] de la méthode TOPSIS.

scénarios / Critère	فَـعَـلَ	فَـعِـلَ	فَـعُـلَ	فَـعِـلٌ	فُـعِـلَ	فَـعَـلُ
juxtaposition des diacritiques	0.13	0.09	0.09	0.04	0.04	0.09
Fréquence d'apparition	0.03	0.008	0.008	0.005	0.008	0.03

Tableau 2: Normalisation de la matrice d'évaluation

- **Pondération de la matrice d'évaluation (normalisée) : e''i,j**
  Cette pondération est faite en appliquant la formule [7] de la méthode TOPSIS.

[6] $e'_{ij} = \frac{f_j(a_i)}{\sqrt{\sum_{t=1}^{m}[f_j(a_i)]^2}}$; i = 1, … , n ; ; j = 1, … , m ; fj (ai) : valeurs déterministes des actions « j » pour le critère « i »

[7] $e''_{ij} = \pi_i \cdot e'_{ij}$, i = 1, …, m ; j = 1, …, n ; avec π_i poids du $i^{ième}$ critère=$\frac{Di}{\sum_1^n Di}$; et $\sum_{i=1}^{n} \pi_{i=1}$

Critère	فَـعَـلَ	فَـعِـلَ	فَـعُـلَ	فَـعِـلٌ	فُـعِـلَ	فَـعَـلُ
juxtaposition des diacritiques	0.30	0.20	0.20	0.10	0.10	0.20
Fréquence d'apparition	0.48	0.15	0.09	0.11	0.11	0.42

Tableau 3: Pondération de la matrice d'évaluation

On utilise pour la pondération des critères l'Entropie (E) [8] et son opposée (D) [9]

Étape 6: Détermination de la solution idéale « e* » et la solution anti idéale « e_*», en adoptant la formule [10] de Topsis

- Critère 1 : « juxtaposition des diacritiques »
  Solution idéale «e_1^* »: 0.33
  Solution anti idéale « e_{1*} »:; 0.33
- Critère 2 : «Fréquence d'apparition»

  Solution idéale «e_2^* » =0. 50 ;
  Solution anti idéale « e_{2*} » = 0.03;

Étape 7: Calculer les mesures d'éloignement pour chaque scénario candidat en utilisant les formules en [11] et [12].

	فَـعَـلَ	فَـعِـلَ	فَـعُـلَ	فَـعِـلٌ	فُـعِـلَ	فَـعَـلُ
D*	0.00	0.34	0.40	0.46	0.46	0.46
D*	0.50	0.15	0.12	0.22	0.22	0.45

Tableau 4: Mesures d'éloignements des scénarios

[8] E_i= -K_i * $\sum_{j=1}^{n}[e'i,j * \log(e'i,j)]$; K_i= constante= 1/(n*log(n))
où n : nombre de scénarios pour le mot ; E_i : entropie du critère i

[9] D_i=1-E_i où D_i est l'opposée de l'entropie ;

[10] $a^* = \{\text{Max } e''_{ij}, i = 1, \dots, m; \text{et } j = 1, \dots, n\}$; $e_j^* = \text{Max}_i\{e''_{ij}\}$.

$a^* = \{ e_j^*, j = 1, \dots, n\} = \{e_1^*, e_2^*, \dots, e_n^*\}$;

$a_* = \{\text{Min } e''_{ij}, i = 1, \dots, m; \text{et } j = 1, \dots, n\}$; $e_{j*} = \text{Max}_i\{e''_{ij}\}$

$a_* = \{ e_{j*}, j = 1, \dots, n\} = \{e_{1*}, e_{2*}\}$;

[11] $D_i^* = \sqrt{\sum_{j=1}^{n}(e''_{ij} - e_j^*)^2}$

[12] $D_{i*} = \sqrt{\sum_{j=1}^{n}(e''_{ij} - e_{j*})^2}$

Étape 8 : Calculer les coefficients des mesures d'éloignement de rapprochement par profil idéal[13].

	فـعـلَ	فـعِـلَ	فـعُـلَ	فـعِـلّ	فـعُـلَ	فـعَـلُ
CR	1	**0.31**	**0.22**	**0.28**	**0.28**	**0.50**

Tableau 5: Coefficients de rapprochement

Étape 9: à partir du tableau précédent, nous allons établir un classement[14] de ces coefficients, selon un ordre décroissant et le scénario ayant obtenu le score le plus élevé sera élu. Dans notre cas ça sera le scénario 1 « فـعـَـلَ », générant ainsi les informations suivantes ;

ذهب	ذَهَبَ	فعل ماض مبني للمعلوم	فَعَلَ	ثلاثي مجرد مسند إلى الغائب(هو) متعد ولازم

Figure 2: Résultat après désambiguïsation par l'analyseur D_Alkhalil

Si le coefficient de rapprochement (CR) est égale à 1, alors un seul scénario est envisageable et l'ambiguïté est levée. Sinon ne seront affichés par D_Alkhalil que les scénarios dont le CR>=0.7. Nous avons choisi un seuil de 70% car après plusieurs essais, nous avons remarqué qu'en deçà de cette valeur, les scénarios corrects pouvaient ne pas s'afficher.

Dans le cas où ce seuil n'est pas atteint, nous affichons les 2 scénarios dont les valeurs sont maximales par rapport à l'ensemble des scénarios proposées pour le même mot.

5 Evaluation d' Alkhalil et D_Alkhalil

Le taux de désambigüisation de l'exemple précédent est à 100% car les verbes sont mieux désambiguïsés par leurs arguments (importance des informations locales), tandis qu'un contexte plus large semble être plus utile pour les noms.

Soit la phrase suivante tirée d'un texte à analyser:

Exemple كتب كثيرة في المكتبة

Les mots de la Phrase à analyser	كتب	كثيرة	المكتبة
Nombre de Scénarios attribués au mot par Alkhalil	17	8	21
Nombre de Scénarios élus pour ce mot par D_Alkhalil	1	1	2
% de Désambiguïsation de notre approche	94.12	87.5	90.47

Taux de désambiguïsation de cette phrase: 90.69% 94%

[13] $C_i^* = \frac{D_{i*}}{D_i^* + D_{i*}}$ i= 1, …, m ; $0 \leq C^*_i \leq 1$.

[14]Rangement des actions suivant leurs ordres de préférences (« i » est meilleur que « j » si $C^*_i > C^*_i$).

On remarque à travers cet exemple que le nombre de scénarios attribués par Alkhalil est de 21 pour le mot « **المكتبة** ». En l'analysant par D_Alkhalil , la première désambigüisation ne lui laisse que le génitif c'est à dire 7 scénarios élus pour 21 probables, la deuxième désambigüisation regroupe les schèmes identiques et applique les étapes de l'AMD pour n'avoir en final que 2 scénarios possibles: الْمَكْتَبَة (el maktabati) et الْمُكَتَّبَة (el moukattabati) qui ne peut les départager qu'une désambigüisation sémantique (Nous utiliserons l'ontologie WordNet arabic [19] avec seulement deux synsets dans ce cas).

Les deux mesures communément utilisées pour évaluer un quelconque système sont: le taux de précision et celui de rappel. Ces deux mesures peuvent être définies par:

$$\text{Précision(P)} = \frac{\text{nombre total de scénarios pertinents retrouvés par le système}}{\text{nombre total de scénarios retrouvés par le système}}$$

Catégorie	Verbes	Noms
Rappel	90,01%	72,71%
Précision	89,76	71.77%

$$\text{Rappel (R)} = \frac{\text{nombre total de scénarios pertinents retrouvés par le système}}{\text{nombre total de scénarios pertinents à retrouver}}$$

Résultats d'une première évaluation

72,71% des noms et 90,01% des verbes deviennent correctement diacrités par notre analyseur D_Alkhalil en tenant compte de l'étiquette grammaticale et en faisant intervenir notre approche de désambigüisation.

6 Conclusion

Le travail que nous avons effectué en désambiguïsation morpho_lexicale nous a d'abord amené à étudier les possibilités d'amélioration de la méthode multi critère d'aide à la décision à base de TOPSIS, qui peut être considérée comme une méthode formelle de désambigüisation. Cette combinaison nous a permis de donner une approche presque réelle en minimisant les résultats de données. Les exemples pris démontrent une désambigüisation d'Al_khalil de plus de 85% pour certains cas.

L'ambiguïté n'a pas été soulevée à son maximum et les résultats obtenus ne sont pas fiables à cent pour cent (100%) car les principales erreurs sont dues à:

- Des données qui ne sont pas répertoriées dans la base de données d' Alkhalil.
- Pour 22,5% des phrases analysées, leur échec d'analyse est dû principalement au fait que leur structure n'est pas couverte par notre grammaire (c'est le cas par exemple de phrases longues ou de phrases anaphoriques et/ou elliptiques non reconnues) ou encore à un échec de segmentation en phrases, échec dans la reconnaissance des caractéristiques morpho_syntaxiques de certains mots, etc.).
- La nécessité d'une autre désambigüisation sémantique que nous envisageons d'implémenter en utilisant l'ontologie de WordNet arabic, qui est précise dans la couverture des sens de termes.

L'analyseur peut être intégré dans d'autres applications et certaines parties de l'analyseur peuvent être réutilisé.

D_Alkhalil n'est qu'à ses débuts, comme extensions en cours:

- Enrichissement de la base de données d'Al_khalil afin de prendre en compte les mots inconnus;
- Étendre les catégories grammaticales d'Alkhalil ;
- Étendre la base des règles de la grammaire de l'analyseur syntaxique.
- Implémentation d'une méthode de désambigüisation sémantique en utilisant le wordnet arabic.

7 Références

[1] ATTIA. M., YASEEN. M. ET CHOUKRI. K., 2005, «Caractéristiques du Corpus arabe écrite produite au sein du projet NEMLAR", www.nemlar.org.

[2] ALLOTTI D., PONSARD C., « Exposé sue les étiqueteurs statistiques et les étiqueteurs par contraintes », 2005.

[3] APIDIANAKI M. « Désambiguïsation Lexicale » Limsi-Cnrs, Groupe Tlp, Orsay, M2r Tal 2011

[4] AUDIBERT L. « Etude Des Critères De Désambiguïsation Sémantique Automatique : Résultats Sur Les Cooccurrences» Taln 2003, Batz-Sur-Mer, 11-14 Juin 2003

[5] AUDIBERT L. « Traitement Automatique Du Langage Naturel, Outils D'analyse De Données Textuelles (Lipn - Umr Cnrs 7030), Université Paris 13 – Laboratoire D'informatique De Paris-Nord (Lipn), 4 Novembre 2010

[6] AUDIBERT L., «Désambiguïsation Lexicale Automatique : Sélection Automatique D'indices. » Taln 2007, Toulouse, 12–15 Juin 2007

[7] BELGUITH_HADRICH L., CHAÂBEN N., ″ Analyse et désambiguïsation morphologiques de textes arabes non voyellés″ TALN 2006, Leuven, 10-13 avril 2006

[8] BOURIGAULT D., Approche Linguistique Pour L'analyse Syntaxique De Corpus. In Cahiers De Grammaire, 25, Université Toulouse Le Mirail (P. Pp.131-151). (2000).

[9] BOURIGAULT D Un analyseur syntaxique opérationnel : SYNTEX, Mémoire présenté pour l'obtention d'une Habilitation à Diriger les Recherches, 2007

[10] DEBILI F., ACHOUR H., SOUSSI E. : « La Langue Arabe Et L'ordinateur : De L'étiquetage Grammatical À La Voyellation Automatique », Correspondances De L'irmc, N° 71, Juillet-Août 2002, Pp 10-28

[11] DEBILI F., SOUISSI E. « Y A-T-Il Une Taille Optimale Des Règles De Succession Intervenant Dans L'étiquetage Grammatical? » Actes De Taln'2005, Dourdan, Juin 2005, 363-372.

[12] DIAB M.: "Automatic Tagging Of Arabic Text : From Raw To Base Phrase Chunks". In Proceedings Of Naacl-Hlt, Boston,Usa,2004

[13] HEINECKE, J., SMITS, G., CHARDENON, C., GUIMIER DE NEEF, E.,MAILLEBUAU, E., & BOUALEM, « Mtilt : Plate-Forme Pour Le Traitement Automatique Des Langues Naturelles ». Traitement Automatique Des Langues, 49(2). (2008).

[14] HOCEINI Y. , ABBAS M. « Méthodologie Multicritère de Désambiguïsation Morphosyntaxique de la Langue Arabe » 3rd International Conference on Arabic Language Processing (CITALA'09), May 4-5, 2009, Rabat, Morocco

[15] HWANG C. R., K. YOON, "Lectures Notes In Economics And Mathematical Systems", Springer-Verlag Berlin Heildelberg, New York, Ny, 1981.

[16] NAVIGLI R. "Word Sense Disambiguation: A Survey". Acm Comput. Surv., 41(2):10 :1–10 :69. (2009).

[17] ROY B., BOUYSSOU D., « Aide Multicritère À La Décision, Méthode Et Cas », Édition: Economica, 1993

[18] SCHARLIG A., « Décider Sur Plusieurs Critères. Panorama De L'aide À La Décision Multicritères », Presses Polytechnique Universitaires Romandes, 1985

[19] TCHECHMEDJIEV A., « État De L'art : Mesures De Similarité Sémantique Locales Et Algorithmes Globaux Pour La Désambiguïsation Lexicale À Base De Connaissances » Actes De La Conférence Conjointe Jep-Taln-Recital 2012, Volume 3: Recital, Pages 295–308, Grenoble, 4 au 8 juin 2012.

[20] TLILI-GUIASSA Y., M.T LASKRI. "Tagging By Combining Rules-Based And Memory-Based Learning", Information Technology Journal 5(4) 2006, Issn1812-5638.

[21] VINCKE P., « L'aide Multicritère À La Décision », Édition: Economica, 1989.

[22] Sourceforge. Opennlp.Maxent. (Http://Maxent.Sourceforge.Net/).(2010).

[23] WOODS W. Transition Network Grammars for Natural Language Analysis, Communications of the ACM, 13, pp. 59-60 (1970),

[24] Rakho M. « Désambiguïsation automatique à partir d'espaces vectoriels multiples clutérisés », rapport intermédiaire université paris 7 – Diderot, juin 2008.

[25] Mars M. « nouvelles ressources et nouvelles pratiques pédagogiques avec les outils tal », 2008