# Evaluation d'une approche possibiliste pour la désambiguïsation des textes arabes (TALN'2014 – Traitement Automatique des Langues Naturelles, Marseille France, 01/07/14-04/07/14)

Raja Ayed, Ibrahim Bounhas, Bilel Elayeb, Narjes Bellamine Ben Saoud, Fabrice Evrard

# Evaluation d'une approche de classification possibiliste pour la désambiguïsation des textes arabes

Raja Ayed[1] Ibrahim Bounhas[2] Bilel Elayeb[1, 3] Narjès Bellamine Ben Saoud[1, 4] Fabrice Evrard[5]

(1) Laboratoire de recherche RIADI, ENSI, Université de la Manouba, 2010, Tunisie
(2) Laboratoire de l'informatique pour les systèmes industriels, Institut Supérieur de Documentation, Université de la Manouba, 2010,Tunisie
(3) Institut de technologies des Émirats, P.O. Box: 41009, Abu Dhabi, Émirats arabes unis
(4) Institut supérieur de l'informatique, ISI, Université de Tunis El Manar, 1002, Tunisie
(5) Institut de recherche en informatique de Toulouse (IRIT), 02 rue Camichel, 31071 Toulouse, France

ayed.raja@gmail.com, bounhas.ibrahim@yahoo.fr, bilel.elayeb@riadi.rnu.tn, narjes.bellamine@ensi.rnu.tn, fabrice.evrard@enseeiht.fr

**Résumé.** La désambiguïsation morphologique d'un mot arabe consiste à identifier l'analyse morphologique appropriée correspondante à ce mot. Dans cet article, nous présentons trois modèles de désambiguïsation morphologique de textes arabes non voyellés basés sur la classification possibiliste. Cette approche traite les données imprécises dans les phases d'apprentissage et de test, étant donné que notre modèle apprend à partir de données non étiquetés. Nous testons notre approche sur deux corpus, à savoir le corpus du Hadith et le Treebank Arabe. Ces corpus contiennent des données de types différents classiques et modernes. Nous comparons nos modèles avec des classifieurs probabilistes et statistiques. Pour ce faire, nous transformons la structure des ensembles d'apprentissage et de test pour remédier au problème d'imperfection des données.

**Abstract.** Morphological disambiguation of Arabic words consists in identifying their appropriate morphological analysis. In this paper, we present three models of morphological disambiguation of non-vocalized Arabic texts based on possibilistic classification. This approach deals with imprecise training and testing datasets, as we learn from untagged texts. We experiment our approach on two corpora i.e. the Hadith corpus and the Arabic Treebank. These corpora contain data of different types: traditional and modern. We compare our models to probabilistic and statistical classifiers. To do this, we transform the structure of the training and the test sets to deal with imprecise data.

**Mots-clés :** Traitement Automatique des Langues Naturelles, Désambiguïsation Morphologique de l'Arabe, Théorie des Possibilités, Classification Possibiliste.

**Keywords:** Natural Language Processing, Arabic Morphological Disambiguation, Possibility Theory, Possibilistic Classification.

## 1 Introduction

De nombreux mots Arabes possèdent la même forme orthographique. Ceci est dû à la richesse morphologique de cette langue (Diab et al., 2004). En effet, l'omission des voyelles courtes peut générer plus de 12 interprétations morphologiques d'un mot donné (Habash et Rambow, 2007). Par conséquent, l'une des formes d'ambiguïté les plus relevées en arabe est l'ambiguïté morphologique. Un mot peut être ambigu à l'égard de sa structure interne. Le traitement morphologique porte sur le morphème qui constitue l'unité élémentaire discernable. L'analyse morphologique d'un mot a pour rôle de déterminer les valeurs d'un grand nombre de caractéristiques ou d'attributs morphologiques d'une entité lexicale (un mot), comme la catégorie grammaticale (nom, verbe, etc.), le genre, le nombre, etc. En fait, un mot non voyellé peut conduire à de nombreuses solutions morphologiques. Par exemple, le mot وقف (wqf), en dehors du contexte, peut être interprété comme وَقَفَ (waqafa, "il s'est levé") ou وَقْفٌ (waqfun, "cession") ou

encore وَقِفْ (waqif, "et lève-toi"), où ce mot est une concaténation de la conjonction و "et" avec le verbe قِفْ "se lever" qui est conjugué à l'impératif. Malgré leur importance, les voyelles courtes ne sont utilisées que dans les textes religieux (Coran, Hadith, etc.) et les manuels didactiques contrairement aux textes modernes trouvés dans les journaux et dans les livres.

L'ambiguïté morphologique se manifeste lorsque l'analyse associe, à une unité lexicale, plusieurs informations non-conformes au contexte du mot, autrement dit quand l'analyse fournit plusieurs valeurs pour certains attributs morphologiques (Hajic, 2000). Par ailleurs, une approche pour la désambiguïsation morphologique arabe est nécessaire pour faire face à l'ambiguïté des mots non voyellés. La désambiguïsation consiste, donc, à attribuer la valeur exacte d'un attribut morphologique parmi celles proposées par l'analyseur. De nombreux travaux utilisent des approches de classification pour résoudre la tâche morphologique de désambiguïsation (Roth et al., 2008).

Nous discutons dans ce papier la contribution d'une nouvelle approche pour la désambiguïsation morphologique arabe basée sur la classification possibiliste. Le but principal est d'apprendre des dépendances morphologiques à partir des textes voyellés et de tester sur des textes non voyellés. Nous organisons ce document comme suit. Tout d'abord, dans la section 2, nous présentons brièvement un état de l'art sur la désambiguïsation morphologique arabe. Quant à la section 3, elle est consacrée à un résumé sur la théorie des possibilités. Notre approche pour la désambiguïsation morphologique possibiliste est détaillée dans la section 4. Les résultats expérimentaux sont présentés et discutés dans la section 5. Nous concluons, dans la section 6 et nous proposons quelques pistes pour de futures recherches.

# 2 La désambiguïsation morphologique arabe

Plusieurs travaux conduisent la désambiguïsation des mots arabes, d'un texte, à l'identification de leurs catégories grammaticales (POS- *part-of-speech*). La désambiguïsation de POS est le fait de déterminer la catégorie grammaticale d'un mot par son utilisation dans un contexte particulier. Elle peut, également, être considérée comme un problème de classification: l'ensemble des valeurs de POS présentent les classes et une méthode de classification est utilisée pour attribuer à chaque occurrence d'un mot (analyse d'un mot) une classe sur la base de la certitude du contexte. L'une des étapes importantes dans la désambiguïsation est la sélection de la méthode de classification. Des méthodes de classification automatique supervisée ont été appliquées. Elles utilisent des techniques d'apprentissage pour apprendre un classifieur à partir des ensembles d'apprentissage annotés (les valeurs de la classe POS sont identifiées). Dans la littérature, les approches de désambiguïsation, se répartissent en trois catégories. Principalement, ces approches sont: les approches à base de règles, les approches statistiques et les approches hybrides qui combinent les deux dernières.

## 2.1 Les approches à base de règles

Les approches à base de règles sont, encore, dites linguistiques. Elles utilisent une base de connaissances des règles écrites par des linguistes permettant d'attribuer des étiquettes aux différentes catégories morphologiques (Daoud, 2009 ; Othman et al., 2004). Nous parlons, principalement, des heuristiques, des règles contextuelles et des règles non contextuelles (Elshafei et al., 2002). Les arbres de décision (Quinlan, 1986) sont conçus pour exposer des bases de règles. Un arbre de décision est un modèle prédictif utilisé pour représenter les règles de classification avec une structure en arbre qui partitionne de façon récursive l'ensemble de données d'apprentissage. Chaque nœud interne d'un arbre de décision représente un test sur une valeur d'un attribut de classification, et chaque branche représente un résultat de test. Une prédiction est faite quand un nœud feuille est atteint. Cette approche est étendue pour extraire et calculer des mesures statistiques utilisées pour l'étiquetage grammatical (Schmid et al., 1994).

## 2.2 Les approches statistiques

Les approches statistiques forment des modèles d'apprentissage à partir des corpus annotés. Elles incorporent des méthodes de classification telles que les modèles de Markov cachés (Garside et Leech, 1987), SVM (Vapnik, 1998), etc. pour calculer des taux de probabilité de chaque valeur résultante d'une catégorie grammaticale d'un mot. Un modèle peut être utilisé pour classer automatiquement les autres textes en se référant aux taux déjà calculés. (Diab et al., 2004) développent un classifieur morphologique utilisant SVM. Ils entrainent et testent le classifieur sur un Treebank arabe de 4000 phrases d'apprentissage et 100 phrases de test. (Habash et Rambow, 2005) utilisent SVM en se basant sur des informations fournies à partir d'un analyseur morphologique. (Mansour et al., 2007) combinent les probabilités calculées sur des ensembles d'apprentissage Arabes et Hébreux pour classer les catégories grammaticales des mots des textes arabes. Ils utilisant les mêmes paramètres de test de (Diab et al., 2004). Quelques travaux de recherches comprennent les modèles de Markov cachés (HMM). (ElHadj et al., 2009) présentent un système d'étiquetage grammaticale qui combine l'analyse morphologique et le modèle de Markov. L'étiqueteur se base sur la structure de la phrase arabe. Dans un premier lieu, le texte est entièrement analysé morphologiquement pour réduire le nombre de valeurs possibles de POS. Dans un second lieu, le modèle statistique (HMM), fondé sur la structure de la phrase arabe,

est utilisé pour attribuer à chaque mot la valeur exacte de sa catégorie grammaticale. (ElHadj et al., 2009) ont utilisé leur propre corpus annoté qui est composé de vieux livres arabes. Le total des mots, dans ce corpus, est environ 21000 mots.

### 2.3 Les approches hybrides

Une approche hybride combine les règles linguistiques avec les informations statistiques afin de résoudre l'ambiguïté morphologique. Dans (Tlili-Guiassa, 2006), on propose une approche qui analyse les affixes grammaticaux et flexionnels et les règles grammaticales en se basant sur l'approche MBL (*Memory based learning*) (Lin et al., 1994). Elle est appliquée pour classer une collection de textes coraniques et éducatifs. (Zribi et al., 2006) combinent l'approche à base de règles avec un étiqueteur trigramme HMM (Collins, 2002). L'apprentissage du classifieur trigramme a été fait sur des textes comportant 6000 mots. Des règles heuristiques ont été appliquées pour sélectionner parmi les résultats proposés.

(Khoja, 2001) a mis en œuvre une approche hybride qui utilise l'algorithme de Viterbi (Forney, 1973; Fettweis et Meyr, 1991). Elle calcule deux probabilités sur un corpus annoté composé de 50000 mots: (i) une probabilité lexicale, qui est la probabilité qu'un mot ait une certaine valeur d'un attribut morphologique spécifié, et (ii) une probabilité contextuelle, qui est la probabilité d'une étiquette à suivre une autre. Une liste de règles grammaticales est préparée à partir de ces statistiques dans le but d'assurer plus de 90% de précision.

Les outils de désambiguïsation linguistiques sont plus rapides et plus efficaces et fiables que les outils statistiques (Hoceini et al., 2011). L'approche linguistique qui n'a besoin que de l'intervention manuelle d'un linguiste, définit un ensemble de règles spécifiques à un domaine particulier. Alors que, les statistiques calculées pour l'apprentissage sont appliquées à n'importe quel domaine de test. Néanmoins, les deux approches statistiques et hybrides nécessitent une phase d'apprentissage dans le but est d'apprendre les paramètres requis pour la désambiguïsation. Par conséquent, l'approche hybride est considérée comme la plus efficace et cohérente en termes d'analyse, car elle combine les deux approches et tire profit de leurs avantages.

La plupart des désambiguïseurs morphologiques arabes ne traitent que la catégorie grammaticale (POS). Les travaux récents (Habash et al., 2009 ; Ayed et al., 2012b) définissent 14 attributs qui décrivent les caractéristiques morphologiques d'un mot. Nous étendons, dans cet article, la classification à ces 14 attributs morphologiques.

## 3 La théorie des possibilités

La théorie des possibilités a été introduite par Zadeh en 1978 pour palier au problème de l'imperfection des données et de l'incomplétude de l'information (Dubois, Prade, 1994). Une information est imparfaite lorsqu'elle est incertaine et/ou imprécise. Nous décrivons, dans les paragraphes suivants, les fonctions, les mesures et les degrés utilisés pour traduire l'incertitude et l'imprécision des données dans la théorie des possibilités.

### 3.1 La distribution de possibilité

La théorie des possibilités est fondée sur la notion de distribution des possibilités désignée par $\pi$. Cette distribution correspond à une application de l'univers de discours $\Omega=\{\omega_1, \omega_2,\ldots, \omega_n\}$ vers l'intervalle [0, 1] modélisant les connaissances du monde réel. Elle distingue les états (les $\omega_i$) plausibles et les états peu plausibles. Les valeurs de cette application sont appelées degrés de possibilités. Si un degré est égal à 1, alors l'état $\omega_i$ associé est plausible. Toutefois, si ce degré est égal à 0 alors l'état est dit impossible.

### 3.2 Les mesures de possibilité et de nécessité

L'imprécision se manifeste quand un état de la réalité est décrit par une variable propositionnelle de valeurs multiples. L'incertitude traduit le fait de ne pas connaître ou prévoir un état de la réalité pour déterminer la valeur de vérité d'une proposition (Dubois et Prade, 1994). Nous évaluons un état par le calcul de deux mesures qui sont, respectivement, la possibilité et la nécessité. Nous désignons *A* un sous-ensemble d'états de l'univers du discours $\Omega$. Nous décrivons la mesure de possibilité de *A*, moyennant une distribution de possibilités $\pi$ (définie sur $\Omega$), comme suit:

$$\Pi(A) = \max_{\omega \in A} \pi(\omega) \quad (1)$$

La mesure de nécessité est extraite à partir de la mesure de possibilité et elle est décrite par:

$$N(A) = \min_{\omega \notin A}[1 - \pi(\omega)] = 1 - \Pi(\bar{A}) \quad (2)$$

Dans la formule 2, $\bar{A}$ définit le complément de A en d'autres termes il englobe les éléments de Ω qui n'appartiennent pas à A. Π(A) évalue le degré de consistance de l'événement *A*. N (A) estime dans quelle mesure *A* est certainement déduit par la connaissance représentée par π. La mesure de nécessité définit le degré auquel on attend l'occurrence d'un événement (Dubois et Prade, 1985).

# 4 L'approche possibiliste de désambiguïsation morphologique

Nous proposons une approche de désambiguïsation morphologique, des textes arabes, basée sur la théorie des possibilités. Plusieurs travaux utilisent les approches de classification pour résoudre l'ambiguïté morphologique (Habash et Rambow, 2005). Un mot est considéré ambigu si l'analyseur morphologique fournit plus qu'une seule solution pour ses attributs morphologiques. La classification assigne une classe à une instance de test donnée. La tâche de désambiguïsation consiste, donc, à accorder à un mot ambigu les valeurs des attributs morphologiques appropriées. Elle est divisée en deux grandes phases qui sont l'apprentissage et le test. Les résultats d'analyse morphologique donnés par les mots voyellés sont, généralement, moins ambigus que ceux donnés par les mots non voyellés. Ainsi, nous proposons d'apprendre à partir des textes voyellés et de tester sur des textes non voyellés.

Pour ce faire, nous commençons par définir l'ensemble d'apprentissage. Cet ensemble est constitué d'une liste d'instances qui sont caractérisées par des attributs avec des valeurs de classes connues. Par conséquent, pour résoudre l'ambiguïté de la catégorie grammaticale (par exemple), nous déterminons d'abord les attributs appropriés qui décrivent chaque instance. En nous inspirant de la technique de classification Yamcha (Diab et al., 2004), nous estimons qu'un attribut morphologique d'un mot est fortement lié à celui des mots qui le précèdent et le suivent. Nous définissons une fenêtre qui contrôle le nombre de mots (avant et après) considérés comme des attributs décrivant la classe d'une instance. Dans des approches existantes, la taille de la fenêtre est 2 (Habash et Rambow, 2005). Notre modèle applique une fenêtre avec une taille quelconque. Pour classer la catégorie grammaticale (POS- *part-of-speech*) d'un mot particulier, si la fenêtre est de 2, nous définissons les attributs POS-2, POS-1, POS+1 et POS+2. Ils indiquent, respectivement, les catégories grammaticales des deux mots précédents et des deux mots suivants. POS peut être décrit par l'ensemble des autres attributs morphologiques, en plus du POS. Nous pouvons utiliser, par exemple, les attributs genre-2, genre-1, nombre+1, nombre+2, etc. La valeur de la classe est la catégorie grammaticale du mot courant. A cet effet, nous identifions, pour chaque mot d'un texte voyellé, 14 attributs morphologiques qui sont *POS, conjonction, particule, déterminant, pronom, personne, voix, aspect, genre, nombre, cas, préposition, mode* et *adjectif*. Ces attributs sont calculés par l'analyseur morphologique Aramorph (Ayed et al., 2012b). Ayant l'exemple de la phrase suivante الرَازِي وَالْبَغْدَادِي دَرَسَا عُلُومَ الطِّبِ (Al-Razi et Al-Bagdadi ont étudié les sciences de la médecine), nous déterminons, l'instance du tableau 1, associée au mot «دَرَسَا (ont étudié)». Pour cette instance, la classe est la catégorie grammaticale (POS) et les attributs utilisés sont les catégories grammaticales des 2 mots adjacents.

| **POS-2** | **POS-1** | **POS+1** | **POS+2** | **POS** |
|---|---|---|---|---|
| NOM_PROPRE<br>الرَازِي (Al-Razi) | NOM_PROPRE<br>وَالْبَغْدَادِي (et Al-Bagdadi) | NOM<br>عُلُومَ (les sciences) | NOM<br>الطِّبِ (de la médecine) | VERBE<br>دَرَسَا (ont étudié) |

TABLEAU 1 : L'instance reliée au mot «دَرَسَا»

L'analyse morphologique d'un mot est fournie indépendamment de son contexte. Dans un texte arabe, même les mots voyellés peuvent donner une analyse morphologique ambigüe. La forme voyellée «إِبْنِ » fournit des valeurs de l'attribut POS à savoir un verbe (tu construis) et un nom (fils de). Par conséquent, les instances d'apprentissage peuvent fournir des informations incomplètes. Ces informations sont dites imprécises lorsque les attributs et ou la classe donnent plus qu'une seule valeur.

Nous pouvons affirmer, clairement, que le contexte nécessaire pour lever l'ambiguïté d'un mot donné est lui-même ambigu ce qui est considérée comme un cas d'imprécision. En effet, la théorie des probabilités est incapable de traiter un tel type de données (imprécises), alors que la théorie des possibilités s'applique naturellement à ces problèmes. Nous proposons des modèles d'apprentissage et de test (classification) basés sur la théorie des possibilités.

### 4.1 L'apprentissage possibiliste des attributs morphologiques

Dans la phase d'apprentissage, nous formons un classificateur pour chaque attribut morphologique. Autrement, nous instaurons un ensemble d'apprentissage pour chaque attribut morphologique. Nous obtenons, globalement, 14 ensembles. Chacun est décrit par les attributs AM ± i où AM forme la totalité des attributs morphologiques et *i* constitue la taille de la fenêtre. Si cette taille est égale à 2, nous obtenons 56 (14x4) attributs d'apprentissage. A chaque mot voyellé est liée une instance décrite par les valeurs de ces 56 attributs et dont la classe est reconnue. Cette classe est l'attribut morphologique associé à l'ensemble d'apprentissage.

Nous devons prendre en compte le fait que les attributs et/ou les classes des instances de classification sont imprécises autrement dit ayant plus qu'une seule valeur possible. L'imprécision est gérée par des distributions de possibilités désignées par π. Soit $T$ un ensemble de données d'apprentissage et $I_k$ l'ensemble des valeurs des attributs de l'instance $k$. On note également $A_j$ le $j^{ième}$ attribut de cet ensemble et $a_{jL}$ une valeur possible de $A_j$. Nous nous inspirons des travaux de Haouari et al. (Haouari et al., 2009) et le modèle de recherche d'information possibiliste développé par Bounhas et al. (2011) pour calculer la fréquence normalisée d'une valeur d'un attribut $a_{jL}$ pour une classe $c_i$ comme suit:

$$Freq\ (a_{jL}, c_i) = \frac{Occ(a_{jL}, c_i)}{Max_{l=1}^{|A_j|} Occ(a_{jl}, c_i)} \quad (3)$$

$Occ(a_{jL}, c_i)$ indique le nombre d'occurrences de la classe $c_i$ avec la valeur $a_{jL}$ c.à.d. le nombre d'instances dont la classe est égale à $c_i$ et la valeur $a_{jL}$ est une valeur possible de l'attribut $A_j$. $|A_j|$ est le nombre de valeurs possibles de $A_j$. Nous utilisons l'opérateur MAX pour obtenir les fréquences normalisées (Bounhas et al., 2011). La somme de toutes les fréquences associées à une classe $c_i$ n'est pas égale à 1 ce qui est l'une des principales hypothèses de la théorie des possibilités afin de traiter des données imparfaites. Dans le cas de l'imperfection des données, le nombre d'occurrences d'une valeur d'un attribut est flou. Nous introduisons une mesure $\beta_{jk}$ appelée le taux de l'imprécision de l'attribut $A_j$ dans l'instance $I_k$ (Haouari et al., 2009). Le nombre d'occurrences est calculé suivant la formule 4 :

$$Occ(a_{jL}, c_i) = \sum_{k=1}^{|T|} \beta_{jk} * \emptyset_{ijkL} \quad (4)$$

Le taux $\beta_{jk} = 1/N$ où $N$ est le produit de $|A_{jk}|$ et $|C_k|$. Ces derniers représentent, respectivement, le nombre de valeurs de $A_j$ dans l'instance $I_k$ et le nombre de classes possibles de $I_k$. Si l'instance est parfaite, alors $\beta_{jk}$ =1. Si dans une instance donnée, un attribut possède deux valeurs et la classe a une seule valeur alors le taux de l'imprécision est égal à 0.5. $\emptyset_{ijkL}$ est égale à 1 si la valeur $a_{jL}$ appartient aux valeurs possibles de $A_j$ dans l'instance $I_k$, et la classe $c_i$ appartient aux valeurs de classes de $I_k$ et 0 sinon.

Les fréquences normalisées sont calculées pour la totalité des instances des différents ensembles d'apprentissage. Elles traduisent les distributions de possibilités de chaque attribut par rapport à une classe.

### 4.2 La classification possibiliste des attributs morphologiques

La classification des 14 attributs morphologiques consiste à désambiguïser chaque mot non voyellé en lui associant les valeurs correctes et précises de ces attributs. Pour ce faire, nous commençons par préparer les instances de l'ensemble de test. En effet, chaque instance décrit un mot non voyellé d'un texte par des attributs de classification qui représentent les mêmes attributs d'apprentissage c.à.d. $AM \pm i$. La classe de l'instance est la valeur correcte à identifier de l'attribut morphologique. Le tableau 2 décrit une instance de test dont l'attribut morphologique à classer est le POS. Pour simplifier la représentation de l'instance, nous nous contentons de 4 attributs de classification à savoir DET-2, POS-1, CONJONCTION-1 et POS+2. Elle est réellement décrite par les 56 attributs. Cette instance est imprécise puisqu'elle donne deux valeurs possibles de l'attribut POS-1.

| DETERMINANT-2 | POS-1 | CONJONCTION-1 | POS+2 | … | POS |
|---|---|---|---|---|---|
| DET | {VERBE; NOM_PROPRE} | NCONJ | NOM | … | ? |

TABLEAU 2 : Un exemple d'une instance de test imprécise

Nous calculons la possibilité de chaque classe $c_i$ par rapport à une instance imparfaite $I_k$ ayant $m$ attributs. Cette mesure s'inspire du classifieur possibiliste de Haouari et al., (2009). La mesure de possibilité est le produit des fréquences de tous les attributs calculées par rapport à l'ensemble d'apprentissage. Cependant, un facteur spécifique est ajouté pour les attributs imprécis. Ce facteur est le taux de l'imprécision $\beta_{jk}$. Par exemple, si un attribut a quatre valeurs possibles, nous calculons le produit des fréquences de ces quatre valeurs et nous introduisons le taux $\beta_{jk}$ égal à ¼. Ainsi, la mesure de possibilité est donnée par la formule 5 :

$$\Pi(c_i|I_k) = \prod_{j=1}^{m} \prod_{L=1}^{|A_{jk}|} Freq\ (a_{jL}, c_i) * \beta_{jk} \quad (5)$$

En se référant à l'instance du tableau 2, si la classe POS possède trois valeurs possibles i.e. NOM, VERBE et NOM_PROPRE, alors trois mesures de possibilités sont à calculer par rapport à cette instance. Ces mesures sont $\Pi(POS = NOM|I_k)$, $\Pi(POS = VERBE|I_k)$ et $\Pi(POS = NOM\_PROPRE|I_k)$. Pour déterminer la mesure $\Pi(POS = NOM|I_k)$, les fréquences nécessaires sont Freq(DETERMINANT-2=DET, POS=NOM), Freq(POS-1=VERBE, POS=NOM), Freq(POS-1=NOM_PROPRE, POS=NOM), etc. Ces fréquences sont calculées dans la phase d'apprentissage.

Un classifieur possibiliste a été défini dans (Ayed et al., 2012a) qui n'évalue pas le pouvoir discriminant des valeurs d'un attribut, car il utilise uniquement la mesure de possibilité (formule 5). Cependant, nous pouvons découvrir que certaines valeurs, d'un attribut donné, ont un plus grand impact dans la résolution de la bonne classe. La théorie des possibilités modélise cet effet par la mesure de nécessité. Elle détermine le degré auquel on attend l'occurrence d'un événement (Elayeb et al, 2009). Cette mesure est donnée par la formule suivante :

$$N(c_i|I_k) = 1 - \prod_{j=1}^{m} \prod_{L=1}^{|A_{jk}|} \left(1 - \frac{\lambda ijL\ * Freq\ (a_{jL}, c_i)}{\beta_{jk}}\right) \quad (6)$$

Où $\boldsymbol{\lambda}_{ijL} = \log_{10}(P/nC_{jL})$

Avec $P$ est le nombre de classes possibles et $nC_{jL}$ est le nombre de classes ayant une fréquence non nulle avec la valeur de la valeur $a_{jL}$ ou en d'autres termes $\text{Freq}\ (a_{jL}, c_i) > 0$.

Nous définissons trois modèles de classification pour déterminer la valeur appropriée d'un attribut morphologique. Le premier modèle se base uniquement sur le calcul des mesures de possibilités. Le deuxième modèle se base sur les mesures de nécessité. Le troisième étant une combinaison des deux, il utilise la somme des mesures de possibilité et de nécessité. La classe choisie correspond à la valeur c*. La meilleure classe de l'instance $I_k$ est celle ayant le plus grand score parmi toutes les classes:

$$c^* = \arg\max_{c_i}(\ \text{score}\ (c_i|I_k) * score(c_i|w_k)) \quad (23)$$

Dans cette formule, score ($c_i$ | $I_k$) peut être égal à $\Pi(c_i|I_k)$ ou $N(c_i|I_k)$ ou $\Pi(c_i|I_k) + N(c_i|I_k)$. Nous introduisons la score lexical score ($c_i|w_k$) (Jurafsky, Martin, 2009). Cette mesure calcule le degré de dépendance d'un mot $w_i$ avec une classe particulière $c_i$ dans l'ensemble d'apprentissage. Si $w_i$ est le mot de l'instance de test $I_k$, alors la possibilité lexicale répond à la question : si nous nous attendions que $c_i$ soit la classe de $I_k$, quelle est la possibilité que le mot soit $w_i$ ? De même ce score peut être calculé de trois manières différentes en utilisant la possibilité et/ou la nécessité.

### 4.3 La classification non possibiliste des attributs morphologiques

Nous visons à comparer les résultats de la classification possibiliste avec les résultats donnés par des classifieurs non possibilistes, afin de désambiguïser les attributs morphologiques. Ces classifieurs ne traitent pas les données imparfaites. Par conséquent, nous proposons de transformer la structuration des données des ensembles d'apprentissage et de test, afin de les préparer pour qu'elles soient utilisées par des classifieurs non possibilistes. Les nouveaux attributs doivent donner des informations précises. Pour ce faire, nous commençons par présenter un ensemble de données imparfaites. La figure 1(a) donne un exemple d'un ensemble d'apprentissage. Nous supposons que la classe à désambiguïser est POS et que les attributs utilisés, pour l'apprentissage et le test, sont POS-1 et CONJUNCTION+1. Cet ensemble est composé de deux instances. La première instance est imprécise, car elle fournit deux valeurs possibles de la classe (NOM_PROPRE et VERBE). La deuxième instance est imprécise puisqu'elle fournit deux valeurs de

l'attribut POS-1 (NOM, VERBE). Nous transformons la structure de données afin d'obtenir un ensemble parfait sans perdre les informations qui s'y trouvent. Pour résoudre le problème de l'imprécision, nous désignons les valeurs, de l'attribut *A*, par $A_i = \{a_1, a_2, ..., a_n\}$. Nous constituons de nouveaux attributs. En effet, nous associons l'attribut *A* à chacune de ses valeurs $a_i$ pour former des attributs notés "*A_a_i*". Ainsi, l'attribut POS-1 a deux valeurs possibles (NOM et VERBE) dans l'ensemble de données de la figure 1. Nous obtenons donc deux attributs POS- 1_NOM et POS-1_VERBE. Nous accordons, aux nouveaux attributs, des valeurs binaires (oui ou non). Pour une instance donnée, si $a_i$ appartient à une des valeurs de l'attribut, *A* alors l'attribut "$A\_a_i$" est égal à "oui". A partir des données de la figure 1 (a), nous formons un nouvel ensemble de données précises (voir figure 1 (b)).

Pour résoudre l'imprécision des classes, nous décomposons une instance en plusieurs ayant chacune une seule valeur de la classe. Si une instance possède *n* valeurs possibles de la classe $\{c_1, c_2, \ldots, c_n\}$, alors nous obtenons *n* instances dont les valeurs des attributs sont similaires. Nous associons à chaque instance une valeur $c_i$.

Les instances dont la classe est précise (ayant une seule valeur) seront dupliquées afin d'augmenter leur poids dans le calcul des mesures de classification. La figure 1(c) présente un ensemble de données parfaites générées à partir des instances de la figure 1(a). Pour lever l'ambiguïté des textes non voyellés moyennant les approches non possibilistes, nous utilisons les méthodes SVM (Vapnik, 1998), le modèle bayésien naïf (Pearl, 1988) et les arbres de décision (Quinlan, 1986). Nous alignons les données au format d'entrée du logiciel WEKA[1]. Ce logiciel fournit des algorithmes d'apprentissage automatique et donne leurs résultats de classification. Nous utilisons WEKA pour classer les attributs morphologiques selon les modèles SVM, les arbres de décision et le modèle bayésien naïf.

| POS-1 | CONJONCTION+1 | POS |
|---|---|---|
| NOM | NCONJ | {NOM_PROPRE; VERBE} |
| {NOM; VERBE} | CONJ | NOM |

(a) Instances Imparfaites

| POS-1_NOM | POS-1_VERBE | CONJONCTION+1_ CONJ | CONJONCTION+1_ NCONJ | POS |
|---|---|---|---|---|
| Oui | Non | Non | Oui | {NOM_PROPRE; VERBE} |
| Oui | Oui | Oui | Non | NOM |

(b) Instances dont les attributs sont précis et les classes sont incertaines

| POS-1_NOM | POS-1_VERBE | CONJONCTION+1_CONJ | CONJONCTION+1_NCONJ | POS |
|---|---|---|---|---|
| Oui | Non | Non | Oui | NOM_PROPRE |
| Oui | Non | Non | Oui | VERBE |
| Oui | Oui | Oui | Non | NOM |
| Oui | Oui | Oui | Non | NOM |

(c) Instances parfaites

FIGURE 1 : Transformation des instances imparfaites en des instances parfaites

# 5 Expérimentations

Dans ce paragraphe, nous décrivons les corpus utilisés pour nos expérimentations. Nous présentons la méthode d'évaluation et les résultats expérimentaux mettant en évidence les aspects de classification possibiliste et non possibiliste.

## 5.1 Les collections de test

L'objectif principal de notre approche est d'acquérir des dépendances morphologiques à partir des textes voyellés et de tester sur des textes non voyellés. En outre, nous considérons les textes arabes classiques, qui ont été ignorés dans des travaux connexes précédents. Par conséquent, nous utilisons une collection d'histoires arabes "Hadiths" qui a fait le

[1] http://weka.wikispaces.com/

sujet de plusieurs travaux (Bounhas et al., 2011 ; Harrag et al., 2013), etc. Les Hadiths parlent de toutes les préoccupations du monde réel et couvrent des connaissances communes et universelles. Pour justifier notre choix, nous estimons que le corpus de hadiths est l'un des rares corpus arabes voyellés. Il contient environ 1400 livres voyellés de hadith, chacun comporte des milliers d'histoires arabes. Les six livres les plus reconnus comprennent plus de 2,5 millions de mots et plus de 95 000 fragments (titres et paragraphes). Par ailleurs, ce corpus est bien structuré et les titres des chapitres et des sous-chapitres représentent des informations contextuelles pertinentes pour désambiguïser des textes (Bounhas et al., 2011). Parmi les textes du corpus de hadiths, nous utilisons six livres encyclopédiques, regroupés par thèmes, qui sont Sahih Al-Bukhari, Sunan Abi Dawud, Sunan Ettermidhi, Sunan Ibn Majah, Sunan Annasaii et Sahih Muslim (Ayed et al., 2012a).

Nous menons nos expérimentations également sur le corpus Arabic Treebank (ATB part 2 v2.0) (Maamouri et al., 2009). Il s'agit d'un corpus de textes arabes non voyellés qui a été produit par *Linguistic Data Consortium*. Ce corpus comprend plus de 500 articles du journal égyptien Al Oumma. Il contient environ 144K de mots annotés (un mot est annoté si on indique la valeur de sa catégorie grammaticale).

Les corpus utilisés présentent deux types de textes i.e. modernes et classiques. Pour pouvoir apprendre les dépendances morphologiques du Hadith, nous passons par l'analyseur morphologique des textes voyellés Aramorph. Cet analyseur nous fournit les valeurs des 14 attributs morphologiques. L'annotation du corpus Arabic Treebank nous donne les valeurs de l'attribut POS. Le test (ou la classification) se fait directement sur les textes non voyellés de Arabic Treebank. Quant aux textes de Hadith, une étape d'élimination des voyelles courtes est indispensable pour pouvoir tester sur des textes non voyellés.

Pour évaluer les résultats des classifications possibilistes et non possibilistes, nous utilisons la méthode de la validation croisée (Kohavi, 1995). En effet, nous formons 10 itérations pour chaque texte du corpus: 90% d'un texte voyellé est utilisé pour l'apprentissage et 10% de mots de ce texte seront classés après avoir éliminé leurs voyelles courtes.

## 5.2 Les résultats expérimentaux

Pour classer les 14 attributs morphologiques, nous procédons comme suit : Tout d'abord nous analysons les textes voyellés de Hadith et nous sauvegardons les solutions morphologiques de chaque attribut. Nous formons, pour tout attribut morphologique $A$, un ensemble d'apprentissage. A chaque mot voyellé est associée une instance. Les instances de cet ensemble sont décrites par les attributs $AM \pm i$ (voir section 4.1) et la classe est l'attribut morphologique $A$. Nous aurons 14 ensembles d'apprentissage. Nous supprimons, par la suite, les voyelles courtes des mêmes textes. Nous formons de la même manière des ensembles de test décrites par les mêmes attributs que les ensembles d'apprentissage. Les valeurs de classes de leurs instances sont non reconnues (ambigües). Elles constituent les attributs morphologiques à classer. Nous désambiguïsons, ensuite, chaque mot de ces textes avec nos trois modèles de classification possibiliste. Pour ce faire, nous calculons les mesures de possibilité et de nécessité en se référant aux fréquences calculées par rapport aux ensembles d'apprentissage (voir section 4). Nous comparons les résultats obtenus avec ceux donnés par les mots voyellés. Pour classer les 14 attributs morphologiques en utilisant les classifieurs non possibilistes, nous utilisons les mêmes structures des instances d'apprentissage et de test.

Les approches non-possibilistes ne supportent pas l'imperfection des données. Nous les transformons en des données parfaites (voir section 4.3) et nous les adaptons au format d'entrée de l'outil Weka pour qu'elles soient appliquées sur des algorithmes de classification de SVM, Arbres de décision et les classifieurs Bayésiens Naïfs. Le tableau 3 présente les taux de désambiguïsation des 14 attributs morphologiques.

Les expérimentations prouvent que les classifieurs possibilistes donnent de meilleurs taux de désambiguïsation par rapport aux classifieurs SVM, Bayésien Naïf et les arbres de décision. Ils en résultent des moyennes de plus de 80% d'instances non voyellés correctement classées. Certains attributs morphologiques donnent les mêmes résultats de classification. Ceci peut être expliqué par le fait que les attributs morphologiques associés fournissent peu de nombres de valeurs de classe (ne dépassant pas 6 chacune). D'un autre côté, l'attribut « PRONOM » (par exemple) offre environ 64 valeurs de la classe qui peut générer des résultats distincts pour les différents classifieurs. Parmi les classifieurs possibilistes, nous remarquons que le modèle qui assemble les mesures de possibilité et de nécessité ($\Pi + N$) fournit de meilleurs résultats (87.43%). Ceci confirme la capacité modèle possibiliste à traiter les données imprécises, surtout que les textes arabes ont un taux d'ambiguïté élevé.

| Attribut morphologique | Classifieur Bayésien Naïf | Arbres de décision | Classifieur SVM | Classifieur possibiliste utilisant *N* | Classifieur possibiliste utilisant Π | Classifieur possibiliste utilisant Π + N |
|---|---|---|---|---|---|---|
| POS | 88.62 % | 89.58 % | 89.98 % | 90.17% | 91.58 % | 90.45% |
| ADJECTIF | 96.51 % | 96.51 % | 96.51 % | 96.86% | 97.58% | 97.63% |
| ASPECT | 71.20% | 71.20% | 71.20% | 86.20% | 81.78% | 86.16% |
| CAS | 56.12 % | 56.12% | 56.12 % | 68.76% | 68.40% | 76.55% |
| CONJONCTION | 83.03 % | 83.03 % | 83.03 % | 88.66% | 95.04% | 90.79% |
| DETERMINANT | 64.12% | 64.16 % | 64.12 % | 95.92% | 95.25% | 96.13% |
| GENRE | 57.15 % | 57.15% | 57.15 % | 90.45% | 93.23% | 93.78% |
| MODE | 99.32 % | 99.32 % | 99.32 % | 99.96% | 99.93% | 99.96% |
| NOMBRE | 85.18 % | 85.18 % | 85.18 % | 87.00% | 95.30% | 93.25% |
| PARTICULE | 96.65 % | 96.65% | 96.65 % | 98.87% | 96.91% | 98.87% |
| PERSONNE | 60.22 % | 60.22 % | 60.22 % | 65.07% | 66.27% | 66.88% |
| PREPOSITION | 82.87 % | 82.87% | 82.87 % | 90.20% | 88.60% | 95.70% |
| VOIX | 71.21 % | 71.21 % | 71.21 % | 78.80% | 78.75% | 79.05% |
| PRONOM | 55.02 % | 55.84 % | 56.88 % | 59.56% | 59.10% | 58.79% |
| **Moyenne** | **76.23 %** | **76.36 %** | **76.46 %** | **85.46%** | **86.27%** | **87.43%** |

TABLEAU 3 : Les taux de désambiguïsation des attributs morphologiques en utilisant 6 classifieurs possibilistes et non-possibilistes dans le corpus du Hadith.

Nous essayons de prouver l'indépendance du domaine de nos modèles possibilistes. Pour ce faire, nous menons nos expérimentations sur le corpus Arabic Treebank rassemblant les textes de journaux. Ce corpus donne les résultats de désambiguïsation de l'attribut POS. A cet effet, les instances des ensembles d'apprentissage et test seront décrites par les attributs POS-2, POS-1, POS+1 et POS+2 qui représentent, respectivement, les catégories grammaticales des deux mots qui suivent et des deux mots qui précédent le mot courant.

Le tableau 4 présente les taux de désambiguïsation de l'attribut POS pour les deux corpus « Hadith » et « Arabic Treebank » données par les six classifieurs.

| | Classifieur Bayésien Naïf | Arbres de décision | Classifieur SVM | Classifieur possibiliste utilisant *N* | Classifieur possibiliste utilisant Π | Classifieur possibiliste utilisant Π + N |
|---|---|---|---|---|---|---|
| **HADITH** | 88.62 % | 89.58 % | 89.98 % | 90.17% | 91.58 % | 90.45% |
| **TREEBANK** | 80.98% | 81.85% | 81.77% | 83.26% | 84.23% | 83.35% |

TABLEAU 4 : Les taux de désambiguïsation de la catégorie grammaticale des deux coprus « Hadith » et «Arabic Treebank»

Nous obtenons des résultats proches avec des taux élevés. Ces résultats révèlent que l'approche de désambiguïsation possibiliste est indépendante du domaine et de type du texte. Elle fournit des taux raisonnables (plus de 80%) pour les textes de journaux ainsi que pour les textes de Hadith. Il y a, cependant, une différence d'environ 7% entre les deux corpus. Comme les tailles des deux corpus sont presque égales, nous pouvons expliquer ce fait par la nature de l'analyseur morphologique (i.e. *Aramorph*) dont le lexique est plutôt classique. Ainsi, cet outil est incapable d'analyser certaines entrées modernes. Par ailleurs, le corpus du Hadith contient des expressions récurrentes, qui existent à la fois dans les ensembles d'apprentissage et de test (par exemple "صلى الله عليه وسلم" ; Paix et la Bénédiction soient Sur Lui).

# Conclusion et perspectives

Nous avons présenté, dans cet article, une nouvelle approche possibiliste pour désambiguïser les attributs morphologiques des textes arabes non voyellés. La désambiguïsation est considérée comme une tâche de classification. A cet égard, nous avons défini un classifieur possibiliste pour apprendre et tester des données imprécises. Nous avons établi trois modèles de classification qui calculent, respectivement, les mesures de possibilité, de nécessité et la somme de ces deux mesures. Nous avons effectué une étude comparative de ces trois modèles de classification possibiliste avec des classifieurs non-possibilistes pour désambiguïser 14 attributs morphologiques. En comparant les résultats des différents classifieurs, nous avons conclu que la théorie possibiliste a donné de meilleurs taux de désambiguïsation quand elle combine les mesures de nécessité et de possibilité.

Malgré ces résultats encourageants, nous avons remarqué que notre approche n'arrive pas à désambiguïser intégralement la totalité des attributs morphologiques. Cela peut s'expliquer par un phénomène linguistique connu en langue Arabe qui se traduit par un ordre relativement aléatoire des mots dans la phrase (Keskes et al., 2013) et également par l'incapacité de désambiguïser les particules qui ont un taux d'ambiguïté élevé, même dans les textes voyellés. Comme perspectives, nous envisageons de faire face à ces problèmes en adoptant l'une des deux alternatives. D'une part, nous pouvons agrandir l'ensemble d'apprentissage. D'autre part, l'intégration d'une analyse linguistique manuelle dans la phase d'apprentissage permettra de filtrer les mots vides et de minimiser le taux d'ambiguïté résultant. Cependant, nous essaierons de réduire le taux d'intervention, pour éviter de traiter tout l'ensemble d'apprentissage à la main. Nous visons aussi à intégrer notre classifieur dans une application de recherche d'information qui traite des textes voyellés et non voyellés, en introduisant une phase primitive de désambiguïsation des requêtes et des documents. A cette étape, nous pouvons renoncer à la désambiguïsation des particules car elles sont considérées comme des mots vides et ne sont pas utilisés pour l'indexation. En outre, les attributs morphologiques calculés par nos outils sont utiles même pour d'autres niveaux d'analyse à savoir syntaxiques et sémantiques (Bounhas et Slimani, 2009).

# Références

AYED R., BOUNHAS I., ELAYEB B., EVRARD F., BENLLAMINE BEN SAOUD N. (2012a). A Possibilistic Approach for the Automatic Morphological Disambiguation of Arabic Texts. *In: T. Hochin & R. Lee (Eds.), Proceedings of the 13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel Distributed Computing (SNPD)*, Kyoto, Japan,187-194.

AYED R., BOUNHAS I., ELAYEB B., EVRARD F., BENLLAMINE BEN SAOUD N. (2012b). Arabic Morphological Analysis and Disambiguation Using a Possibilistic Classifier. *In Intelligent Computing Theories and Applications, Proceedings of the 8th International Conference on Intelligent Computing (ICIC)*, China, 274-279.

BOUNHAS I., SLIMANI Y. (2009). A hybrid approach for Arabic multi-word term extraction. *In Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE)*, Dalian, China, 1–8.

BOUNHAS I., ELAYEB B., EVRARD F., SLIMANI Y. (2011). Organizing Contextual Knowledge for Arabic Text Disambiguation and Terminology Extraction. *Knowledge Organization* 38(6):473–490.

COLLINS M. (2002). Discriminative training methods for hidden Markov models: theory and experiments with perceptron algorithms. *In Proceedings of the ACL-2nd conference on Empirical methods in natural language processing*, Stroudsburg, PA, USA, 1-8.

DAOUD D. (2009). Synchronized Morphological and Syntactic Disambiguation for Arabic. *Advances in Computational Linguistics* 41, 73-86.

DIAB M., HACIOGLU K., JURAFSKY D. (2004). Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks. *In Human Language Technology conference / North American chapter of the Association for Computational Linguistics annual meeting*, Boston, USA, 149-152.

DUBOID D., PRADE H. (1985). Théorie des possibilités: applications à la représentation des connaissances en informatique. Masson, Paris, France.

DUBOID D., PRADE H. (1994). Possibility Theory: An Approach to computerized Processing of Uncertainty. Plenum Press, New York, USA.

ELAYEB B., EVRARD F., ZAGHDOUD M., BEN AHMED M. (2009). Towards an intelligent possibilistic web information retrieval using multiagent system. *Interactive Technology and Smart Education* 6(1): 40–59.

ELHADJ Y., AL-SUGHAYEIR I., AL-ANSARI A. (2009). Arabic Part-Of-Speech Tagging using the Sentence Structure. *In: Proceedings of the Second International Conference on Arabic Language Resources and Tools*, Cairo, Egypt, 241-245.

ELSHAFEI M., AL-MUHTASEB H., AL-GHAMDI M. (2002). Techniques for high quality Arabic speech synthesis. *Information Sciences 140*(3), 255-267.

FETTWEIS G., MEYR H. (1991). High-speed parallel Viterbi decoding: algorithm and VLSI-architecture. *IEEE Communications Magazine*, 46- 55.

FORNEY G.D. (1973). The Viterbi algorithm. *Proceedings of IEEE* 61: 268-278.

GARSIDE R., LEECH F. (1987). The UCREL probabilistic parsing System. *The Computational Analysis of English: A Corpus-Based Approach*, Longman, London, 66-81.

HABASH N., RAMBOW O. (2005). Arabic Tokenization, Part-of-speech Tagging and Morphological Disambiguation in One Fell Swoop. *In: the proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Stroudsburg*, PA, USA, 573–580.

HABASH N., RAMBOW O. (2007). Arabic Diacritization Through Full Morphological Tagging. *In: Human Language Technologies: The Conference of the North American Chapter of the Association for Computational Linguistics, Stroudsburg*, PA, USA, 53-56.

HABASH N., RAMBOW O., ROTH R. (2009). Mada+tokan: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization. *In Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR).*

HAJIC J. (2000). Morphological Tagging: Data vs. Dictionaries. *In: Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference, Stroudsburg*, PA, USA, 94-101.

HAOUARI B., BEN AMOR N., ELOUEDI Z., MELLOULI K. (2009). Naïve possibilistic network classifiers. *Fuzzy Sets and Systems* 160(22): 3224-3238.

HARRAG F., ALOTHAIM A., ABANMY A., ALOMAIGAN F., ALSALEHI S. (2013). Ontology Extraction Approach for Prophetic Narration (Hadith) using Association Rules. *International Journal on Islamic Applications in Computer Science And Technology* 1(2): 48-57.

HOCEINI Y., CHERAGUI M. A., ABBAS M. (2011). Towards a New Approach for Disambiguation in NLP by Multiple Criterian Decision-Aid. *The Prague Bulletin of Mathematical Linguistics 95*, 19-32.

JURAFSKY D., MARTIN J.H. (2009). Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition. *Pearson Prentice Hall, Upper Saddle River*, New Jersey, USA.

KESKES I., BEANAMARA F., HADRICH BELGUITH L. (2013). Segmentation de textes arabes en unités discursives minimales. *TALN-RECITAL, Les sables d'Olonne,* 435-449.

KHOJA SH. (2001). APT: Arabic part-of-speech tagger. *In: Proceedings of Student Workshop at the Second Meeting of the North American Association for Computational Linguistics*, Carnegie Mellon University, Pennsylvania, USA.

KOHAVI R. (1995). A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection. *In: Proceedings of the 14th International Joint Conference on Artificial Intelligence,* San Francisco, CA, USA, 1137-1143.

LIN J., VITTER S, J., HELLERSTEIN L. (1994). A Theory for Memory-Based Learning. *Machine Learning* 17(2-3): 143-167.

MAAMOURI M., BIES A., KULICK S. (2009). Creating a Methodology for Large-Scale Correction of Treebank Annotation: The Case of the Arabic Treebank. *In the proceedings of MEDAR Second International Conference on Arabic Language Resources and Tools*, Cairo, Egypt, 138–144.

MANSOUR S., SIMA'AN K., WINTER Y. (2007). Smoothing a lexicon-based pos tagger for Arabic and Hebrew. *ACL07 Workshop on Computational Approaches to Semitic Languages*, Prague, Czech, 97-103.

OTHMAN E., SHAALAN K., RAFEA A. (2004). Towards Resolving Ambiguity in Understanding Arabic Sentence. *In the proceedings of International Conference on Arabic Language Resources and Tools, NEMLAR*, Egypt, 118-122.

PEARL J. (1988). Probabilistic reasoning in intelligent systems: networks of plausible inference. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmman, San Francisco, California, USA.

QUINLAN J. R. (1986). Induction of decision trees. *Machine Learning* 1(1): 81-106.

ROTH R., RAMBOW O., HABASH N., DIAB M., RUDIN C. (2008). Arabic Morphological Tagging, Diacritization, and Lemmatization Using Lexeme Models and Feature Ranking. *In: Proceedings of the Association for Computational Linguistics conference (ACL)*, Columbus, Ohio, USA, 117-120.

SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. *In Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK, 44-49.

TLILI-GUIASSA Y. (2006). Hybrid Method for Tagging Arabic Text. *Journal of Computer Science* 2(3): 245-248.

VAPNIK V. (1998). Statistical Learning Theory. Wiley, New York, USA, 1-736.

ZRIBI C., TORJMEN A., BEN AHMED M. (2006). An Efficient Multi-agent System Combining POS-Taggers for Arabic Texts. *In Proceedings of 7th international conference of Computational Linguistics and Intelligent Text Processing*, LNCS Volume 3878, Springer, 121-131.