Une Analyse Morphologique de la Langue Arabe basée sur l'Aide Multicritère à la Décision

CHERAGUI Mohamed Amine
Université d'Adrar
Adrar, Algérie
m_cheragui@univ-adrar.dz

***Résumé*—Dans cet article, nous présentons nos travaux sur la morphologie arabe et plus particulièrement les mécanismes de la levée de l'ambiguïté morphologique dans le texte arabe. Ces recherches qui ont donnés naissance au système TAGHIT qui est un étiqueteur morphosyntaxique de la langue arabe, où l'originalité de notre travail réside dans l'implémentation à l'intérieur de notre système d'une nouvelle approche de désambiguïsation différente de celles qui existent actuellement (contraintes et stochastique) qui se base sur les principes et techniques issues de l'aide multicritère à la décision.**

Mots clés-TALN[1]; TALA[2]; Ambiguité morphologique; étiquetage morphosyntaxique ; AMD[3].

I. INTRODUCTION

Si le traitement automatique de la langue arabe a connu ces dernières décennies une véritable ascension que ce soit sur le plan scientifique ou socio-économique, et cela, par la création d'entreprises mais aussi par l'émergence d'un nombre important de produits spécialisés, comme : les traducteurs automatiques, correcteurs orthographiques d'erreurs, générateurs automatiques de résumés, …etc. Il reste tout de même confronté à un problème qui le pénalise dans son processus de développement, à savoir l'ambiguïté. L'ambiguïté est considérée aujourd'hui comme étant la principale pierre d'achoppement du traitement automatique des langues à une époque où la mémoire de stockage et la puissance de traitement des ordinateurs ne constituent plus un frein au développement d'applications informatiques [3]. Par définition l'ambiguïté est comparée à un état de confusion, cet embrouillement se manifeste sous différentes formes est selon les différents niveaux de traitement que ce soient : lexical, morphologique, syntaxique et même sémantique, induisant le plus souvent à des résultats d'analyses erronées dues à la présence de plusieurs solutions. Dans cet article, nous nous intéresserons à l'ambiguïté morphologique dans le traitement automatique de la langue arabe. En donnant en premier lieu une description complète de ce phénomène. Ensuite, nous exposons les différentes approches qui ont été mis en place pour lever cette forme d'ambiguïté. Après, Nous présenterons notre approche de désambiguïsation morphologique basée sur l'aide multicritère à la décision qui est implémentée à travers le système TAGHIT : un étiqueteur morphosyntaxique de la langue arabe.

II. LA MORPHOLOGIE: ANALYSE ET AMBIGUÏTE

La morphologie est un domaine de la langue qui permet la description des règles régissant la structure interne des mots (unités lexicales), chez les grammairiens la morphologie est l'étude des formes des mots (flexion et dérivation), en d'autres termes, la morphologie est l'étude des mots considérés isolement (hors contexte) sous le double aspect de la nature et les variations qu'ils peuvent subir. En langue arabe, l'analyse morphologique est d'autant plus importante que les mots sont fortement agglutinés, c'est-à-dire qu'ils sont formés dans leur majorité par assemblage d'unités lexicales et grammaticales élémentaires. Ainsi le principal objectif d'une analyse morphologique automatique et de reconnaitre ces unités et d'attribuer à chacune divers types d'informations telles que la catégorie grammaticale et les traits morphologiques (genre, nombre, la voix, le mode, …etc.). C'est à ce stade, que l'ambiguïté morphologique se manifeste le plus souvent; lorsque l'analyse assigne à une unité lexicale plusieurs informations ce qui génère la notion de combinatoire [10]. Prenant comme exemple, le mot «أحمدُ », ce dernier peut prendre plusieurs interprétations (voir figure 1) et cela est dû au contexte.

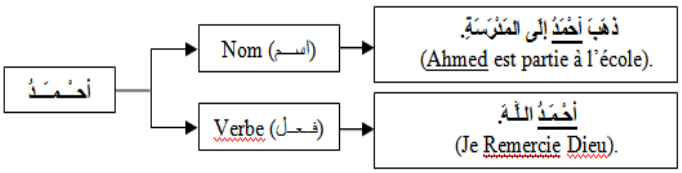

Figure 1. Exemple d'une ambiguïté morphologique.

III. ETIQUETAGE MORPHOSYNTAXIQUE POUR LEVER L'AMBIGUÏTE MORPHOLOGIQUE

A. *Qu'est ce que le Tagging?*

Le Tagging (étiquetage ou marquage) est le fait d'assigner une étiquette ou un tag à un mot. Le tag contient des informations morphosyntaxiques sur le mot, c'est-à-dire des informations concernant la forme et la fonction du mot. Elles comprennent notamment la catégorie grammaticale, le genre, le nombre, le temps et le mode [5].

1 TALN : Traitement Automatique des Langues Naturelles.
2 TALA : Traitement Automatique de la Langue Arabe.
3 AMD : Aide Multicritère à la Décision.

B. Principe de Base

Pour mettre en place un étiqueteur morphosyntaxique de la langue arabe, on doit construire trois (03) modules qui seront complémentaires, ces modules sont :

- *Module de segmentation :* lorsqu'on évoque la segmentation dans le traitement automatique des langues dites naturelles, on parle le plus souvent de trois (03) niveaux de segmentations :
  - Segmentation au niveau du texte ;
  - Segmentation au niveau de la phrase ;
  - Segmentation au niveau du mot.
- Module d'analyse morphologique : Le but principal de ce type d'analyse est de vérifier l'appartenance d'un mot donné au domaine linguistique choisi et de pouvoir disposer ainsi de tous les renseignements le concernant pouvant servir à l'analyse syntaxique.
- Module de désambiguïsation : La désambiguïsation est une étape cruciale dans le processus d'étiquetage morphosyntaxique, à ce niveau du traitement si un mot est mal étiqueté, les règles de la grammaire s'appliqueront mal ou pas du tout. Cependant la phase de désambiguïsation n'est pas toujours nécessaire ou obligatoire au bon déroulement du processus d'étiquetage. Il faut dire que le module de désambiguïsation rentre en jeu dans un seul cas de figure, celui où l'unité lexicale (mot) reçoit plus d'une étiquette (plus d'une information morphosyntaxique), ce qui va générer une situation de confusion ou ambiguïté. C'est à ce stade que notre contribution va apparaître en présentant une nouvelle démarche de désambiguïsation différente de celle qui existe actuellement basée sur l'aide multicritère à la décision.

IV. LEVEE L'AMBIGUÏTE MORPHOLOGIQUE

La désambiguïsation morphologique a fait l'objet de plusieurs travaux de recherches qui sont classés principalement en deux (02) approches et chaque approche englobe une ou plusieurs techniques [2], [4], [1], [11], [5] :

A. L'approche par contraintes

L'approche de désambiguïsation par contraintes (ou par règles) est la méthode la plus ancienne qui a été mis en place pour remédier au problème de l'ambiguïté morphologique. Cette approche se base principalement sur l'intervention d'un linguiste (ou un grammairien) afin d'établir une liste de règles permettant de lever l'ambiguïté. Ces règles sont généralement classées en catégories de type: grammatical, structural, sémantique, logique, …etc [1], [11].

B. L'approche Stochastique (Statistique / Probabiliste)

Levée l'ambiguïté morphologique en adoptant une approche stochastique, celà consiste essentiellement à définir deux (02) sortes d'informations, la première est sur le mot à étiqueter (i.e. l'association entre le mot et l'étiquette) et la deuxième information est contextuelle syntaxique (i.e. la possibilité de déterminer la probabilité d'avoir une étiquette « i » quant elle est précédée de l'étiquette « j » dans le texte). En plus de ces deux (02) hypothèses qui sont considérées comme étant les formules de calcul, une phase d'apprentissage est obligatoire, et celà, en entraînant le module de désambiguïsation sur un corpus généralement annoté (i.e. étiqueter à la main) au préalable [5].

V. UNE APPROCHE DE DESAMBIGUISATION BASEE SUR L'AMD

C'est à ce stade que notre contribution va apparaître, par le fait de proposer une nouvelle approche de désambiguïsation qui se base sur un formalisme purement mathématique issue de la méthodologie multicritère.

A. Démarche

Adopté une approche AMD, celà consiste à suivre une certaine démarche qui se résume dans les points suivants [8]:

- Etape 1 : *Dresser la liste des scénarios candidats.* La mise en place d'un ensemble « A » qui contient tous les scénarios possibles, dans notre cas il s'agit des étiquettes ambiguës « a_i ».
- Etape 2 : *Construire une famille cohérente de critères.* Une bonne application d'une démarche multicritère passe impérativement par un bon choix concernant les critères sur lesquels le calcul sera posé. Ces critères seront définis en se basant sur les notions : de cohérence, d'indifférence, de préférence stricte ou faible ou de non comparabilité [9].
- Etape 3 : *Définir une fonction d'évaluation pour chaque critère*. Cette fonction doit être maximisée ou minimisée suivant la nature du critère définit.
- Etape 4 : *Etablir le tableau (ou matrice) d'évaluation.* Les résultats des évaluations de chaque scénario candidat suivant tous les critères sont groupés dans un tableau d'évaluation.
- Etape 5 : *L'agrégation et la pondération*.
  1. **L'agrégation :** Le but est de réduire le nombre de scénarios. Le choix d'une méthode d'agrégation va permettre de normaliser et pondérer le tableau d'évaluation ainsi que de classer les scénarios selon leurs scores globaux. Afin d'agréger les différentes évaluations d'un scénario calculées selon les critères retenus, nous proposons d'appliquer **la méthode TOPSIS**[5].
  2. **La pondération :** consiste à déterminer le poids de chaque critère selon son importance. Pour pondérer les différents critères nous adoptons **la méthode Entropie**.

B. La méthoded'agrégation TOPSIS

Créée par Hwang et Yoon en 1981 [6], cette méthode se base sur la relation de dominance qui résulte de la distance par rapport à la solution idéale. Son fondement consiste à choisir

4 Processus d'ajout d'affixes à un mot qui exprime ses différentes relations grammaticales.

5 TOPSIS: Technique for Order by Similarity to Ideal Solution.

une solution qui se rapproche le plus de la solution idéal et de s'éloigner le plus possible de la pire solution qui dégrade tous les critères. Ainsi, nous calculons pour chaque scénario de désambiguïsation « a_i » un score d'évaluation global « $C(a_i)$ » qui représente la somme pondérée des différentes évaluations de « a_i » selon tous les critère de « F ». Cette méthode est composée de six (06) phases

- Phase 1: Calcul du tableau d'évaluation normalisé :

$$e'_{ij} = \frac{g_j(a_i)}{\sqrt{\sum_{i}^{m} [g_j(a_i)]^2}} \quad \text{Avec: } \begin{matrix} i=1,\ldots,m. \\ j=1,\ldots,n. \end{matrix} \quad \textbf{(1)}$$

Où :

Les « $g_i(a_i)$ » correspondent aux valeurs déterministes des scénarios « i » pour le critère « j ».

- Phase 2 : pondération du tableau d'évaluation normalisée :

$$e''_{ij} = \pi_j \cdot e'_{ij} \quad \text{Avec: } \begin{matrix} i=1,\ldots,m. \\ j=1,\ldots,n. \end{matrix} \quad \textbf{(2)}$$

Où :

«π_j» est le poids du $j^{ième}$ critère

- Phase 3 : Détermination de la solution idéale (a^+) et de la solution anti-idéale (a_-) par rapport à chaque critère.

$$a^+ = \{\max_i e''_{ij},\ i=1,\ldots,m ;\ \text{et } j=1,\ldots,n\} ; \quad \textbf{(3.1)}$$

$$a_- = \{\min_i e''_{ij},\ i=1,\ldots,m ;\ \text{et } j=1,\ldots,n\} ; \quad \textbf{(3.2)}$$

- Phase 4 : Calcul des mesures d'éloignements (*i.e. Calculer la distance euclidienne par rapport aux profils a^+ et a_-*) :

$$D_{i-} = \sqrt{\sum_{j}^{n} (e''_{ij} - e_{j-})^2}. \quad \text{Avec : } i=1,2,\ldots,m. \quad \textbf{(4.1)}$$

$$D_i^+ = \sqrt{\sum_{j}^{n} (e''_{ij} - e_j^+)^2}. \quad \text{Avec : } i=1,2,\ldots,m. \quad \textbf{(4.2)}$$

- Phase 5 : Calculer les coefficients de rapprochement au profil idéal :

$$C_i^+ = \frac{D_{i-}}{D_i^+ + D_{i-}}. \quad \text{Avec: } \begin{matrix} i=1,\ldots,m ; \\ 0 \le C_i^+ \le 1. \end{matrix} \quad \textbf{(5)}$$

- Phase 6 : Rangement des scénarios suivant leur ordre de préférences (*i.e. en fonction des valeurs décroissantes de C_i^+ ; « i » est meilleur que «j » si $C_i^+ > C_j^+$*).

C. *Laméthode de pondération Entropie*

L'Entropie est une méthode de pondération objective qui permet d'attribuer des poids aux critères en se basant sur le facteur de discrimination (*i.e. plus le critère est important plus il aura un poids élevé*) [7]. Les étapes de cette méthode sont données comme suites :

- Etape 1 : L'entropie d'un critère « j » est calculée par la formule [7]:

$$E_j = -K \sum_{i}^{n} g_j(a_i) \log[(g_j(a_i))]. \quad \textbf{(6)}$$

Où :

K : est constante choisie de telle sorte que pour tous « j », on a $0 \le E_j \le 1$, pour notre cas « K » est calculé suivant la formule :

$$K = \frac{1}{\text{Log}(n)}. \quad \textbf{(7)}$$

Où :

n : le nombre de scénarios de désambiguïsation.

- Etape 2 : Les poids seront calculés en fonction de la mesure de dispersion (opposée de l'entropie) :

$$D_j = 1 - E_j. \quad \text{Avec: } j=1,\ldots,n. \quad \textbf{(8)}$$

- Etape 3 : Les poids seront ensuite normalisés par :

$$W_j = D_j / \sum_{j}^{n} D_j \quad \text{Avec: } j=1,\ldots,n. \quad \textbf{(9)}$$

VI. PRÉSENTATION DU SYSTÈME TAGHIT

Pour appliquer notre approche de désambiguïsation, nous avons conçu tout un étiqueteur morphosyntaxique dédie au texte arabe que nous avons baptisé TAGHIT (TAGger morpHsyntaxique InformaTisé) qui a été développé en langage Python. La figure 2 présente l'architecture générale de notre système TAGHIT.

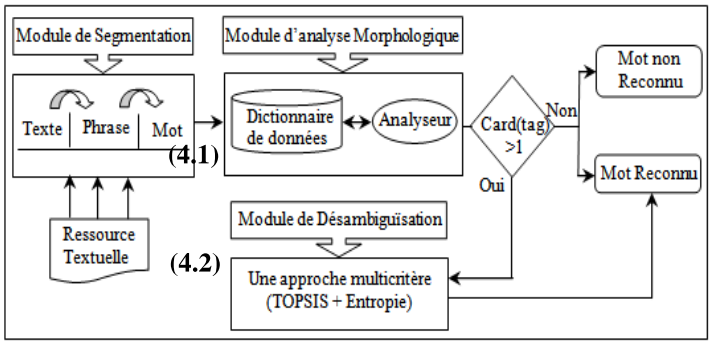

Figure 2. Architecture générale du système TAGHIT.

Comme on peut le constater sur la figure le système TAGHIT est composé de trois (03) modules, complémentaire, qui sont :

- Module de segmentation (texte, phrase, mot);
- Module d'analyse morphologique ;
- Module de désambiguïsation.

VII. EXEMPLE ILLUSTRATIF

Soit la pharse Ph= « رَجَـعَ المـُغـْتَـرِبُ إلى الـوَطـَنِ » (Revenu l'imigrant au pays)[6], qui se trouve à l'entré de notre étiqueteur. Après segmentation de la phrase en unités, l'analyse se fait sans problème pour l'unité 2 (المـُغـْتَـرِبُ), 3 (إلى), et 4 (الـوَطـَنِ),

cependant l'unité 1 « رَجَـعَ » présente un cas typique d'ambiguïté morphologique. Pour lever cette ambiguïté nous allons appliquer notre approche de désambiguïsation multicritère, selon la démarche suivante :

- ***Etape 1 :*** **Construire la liste des scénarios d'analyse.**
  Cette liste est obtenue directement après le processus d'analyse morphologique. Ce qui va générer l'ensemble « A ».

Verbe	Scénario (Schème)	langueur
رجـع	فـَعـَلَ	6
	فـَعـِلَ	6
	فـَعـُلَ	6
	فـُعـِلَ	6

Tableau 1. Exemple d'ambiguïté générée lors de l'analyse du verbe « رجَـعَ ».

- ***Etape 2 :*** **Définition des critères et des fonctions dévaluations.**
  Afin de construire une famille cohérente de critères « F », nous proposons deux critères de base pour discriminer entre les scénarios d'analyse. Ces critères sont :
  - **Critère 1 : Concordance des voyelles**

  Ce critère va vérifier la concordance entre les voyelles de l'unité lexicale et les voyelles de chaque scénario candidat, de telle sorte que chaque concordance vaut : un (1). La fonction d'évaluation est l'addition (+).
  - **Critère 2 : La Fréquence d'apparition**

  Ce critère s'appui sur un calcul statistique fait sur la base d'un corpus annoté, de telle manière que le scénario qui se manifeste le plus souvent (une plus grande fréquence d'apparition) aura systématiquement le score le plus élevé. Chaque apparition vaut : un (1). La fonction d'évaluation est l'addition (+).
  Remarque :
  Le corpus utilisé est composé de plus de 300 unités réparties sur 10 paragraphes choisis arbitrairement à partir des livres scolaires de l'école algérienne.
- ***Etape 4 :*** **Générer le tableau (matrice) d'évaluation**

	S1	S2	S3	S4
Critère 1	3	2	2	1
Critère 2	16	5	2	1

Tableau 2. Tableau (Matrice) d'évaluation.

- **Etape 5: Application de la méthode d'agrégation et de pondération**. Les résultats de l'application de la méthode d'agrégation TOPSIS et de pondération Entropie sont donnés par le tableau 3.

	S1	S2	S3	S4
C^+	1	0.32	0.24	0

Tableau 3. Mesures d'éloignements.

- **Etape 6 : Etablir un classement décroissant des coefficients.** le scénario ayant obtenu le score le plus élevé sera élu.

  $C_1^+=1 > C_2^+=0,32 > C_3^+=0,24 > C_4^+=0.$

  Selon notre approche le scénario 1 « فـَعـَلَ » sera sélectionné par le système, ainsi les informations morphologiques suivantes seront générées.

	Information
Verbe	رَجَـعَ
Schème	فـَعـَلَ
Etiquette	VAA3PMSIA
Désignation en français	Verbe Accompli Actif 3e Personne Masculin Singulier Invariable Accusatif.
Désignation en arabe	فعل ماضي مبني للمعلوم للمفرد المذكر الغائب، مبني على الفتح.
Racine	رجـع

Tableau 4. Informations générées de l'étiquetage du verbe « رَجَـعَ ».

VIII. CONCLUSION

Nous avons pu à travers cet article présenté notre système TAGHIT dédié à l'étiquetage morphosyntaxique du texte arabe. L'originalité dans notre système est qu'il est doté d'une nouvelle approche de désambiguïsation basée sur l'AMD, permettant un classement multicritère des scénarios de désambiguïsation. Cette technique bien quelle soit peu exploité montre que la cohabitation entre une démarche multicritère et le traitement automatique des langues naturelles est à la fois possible mais aussi avantageuse. Cette nouvelle approche offre une alternative de choix par rapport à l'approche stochastique et peut être complémentaire avec l'approche par contraintes.

Comme l'ambiguïté morphologique dans le texte arabe se manifeste principalement au niveau du verbe et du nom dérivable, la première version de notre étiqueteur a été limitée dans la phase de désambiguïsation aux verbes, les premiers résultats obtenus sont encourageants. Nous somme actuellement entrain de définir d'autres critères d'évaluations touchant l'aspect grammatical (la position du mot dans la phrase). Ainsi notre système sera plus fiable et robuste.

REFERENCES

[1] A. Abeillé, P. Blache, "Grammaires et analyseurs syntaxiques. Edition Ingénierie des langues," Paris, Hermès 2000.

[2] D. Alloti, C. Ponsard, "Exposé sur les étiqueteurs statistiques et étiqueteurs par contraintes," 2005.

[3] P. Boulillon, F. Vandooren, L.. Da Sylva, L. Jacqmin, S. Lehmann, G. Russell et E. Viegas, "Traitement automatique des langues naturelles," Champs linguistique, Edition Duculot 2005.

[4] J. P. Chanod, P. Tapanainen, "Les étiqueteurs statistiques et les étiqueteurs par contraintes," 1995.

[5] B. Merialdo, "Tagging english text with a probabilistic model," Computational linguistics, 1994.

[6] C. R. Hawang, K. Yoon, " Lecture Notes in Economics and Mathmatical system," Springer-Verlag Berlin Heidelberg, NewYork 1981.

[7] J. C. Pomerol, S. B. Romero, "Choix multicritère dans l'entreprise: principes et pratique," Edition Hermès 1993.

6 La traduction est faite mot par mot.

[8] P.Vinck, "L'aide multicritère à la décision. SMA," université de Bruxelles, 1989.

[9] B. Roy, "Méthodologie multicritères d'aide à la décision," Edition ECONOMICA 1993.

[10] J. Vergne, E. Giguet, "Regards théorique sur le Tagging," Proceedings of thefifth annual conference Le Traitement Automatique des Langues Naturelles 1998.

[11] L. Al-Sulaiti, "Deigning and Developing a Corpus of Contemporary Arabic," School of Computing, University of Leeds (UK), March 2004.