

Méthodologie Multicritère de Désambiguïsation Morphosyntaxique de la Langue Arabe

HOCEINI Youssef * and ABBAS Moncef **

* Institut d'informatique, Université de Bechar, Bechar ; Algérie, y_hoceini@yahoo.fr

** USTHB, Faculté de Mathématiques Laboratoire LAID3 , Dépt. R.O., Alger, Algérie, moncef_abbas@yahoo.com

Résumé -- L'un des problèmes qui défient des chercheurs travaillant sur le TALN¹ est sans doute les ambiguïtés générées par les applications linguistiques quelles soit morphologiques, syntaxiques ou sémantiques. De ce fait, une multitude de scénarios de résultats est à prendre en considération. En effet, une confusion détectée entre deux ou plusieurs alternatives pendant une phase d'analyse engendre au moins deux solutions concurrentes. Le problème majeur dans un tel cas réside dans le choix du meilleur scénario possible.

Le contexte linguistique précis de l'arabe insiste sur la présence d'une multitude de critères qui témoignent de la fonction de plusieurs contraintes (grammaticale, sémantique, logique et statistiques). Dans ce domaine, la réalisation d'un système performant exige qu'il soit robuste, rapide et moins ambigu afin de permettre une analyse grammaticale correcte.

Le but de cet article, outre de faire une description du phénomène de l'ambiguïté et les approches existantes de la levée d'ambiguïté telles que l'utilisation des modèles mathématiques et stochastiques (probabiliste et statistique) ou l'introduction des modèles par contraintes -à base de règles-, vise à analyser et modéliser un prototype d'un système de désambiguïsation morphosyntaxique de l'arabe en proposant une méthode originale basée sur la théorie décisionnelle sous une approche AMD² permettant le classement multicritère des scénarios de désambiguïsation en vue d'en faire émerger le meilleur. Cette démarche a l'avantage de réduire les scénarios dominés et de classer le reste selon différents critères d'évaluation.

Mots clés -- TALN ; AMD ; ambiguïté ; désambiguïsation ; système morphologique de l'arabe ; analyse morphosyntaxique ; étiquetage ;

modèle probabilistes ; modèle par contraintes ; classement multicritère

I. INTRODUCTION

Si jadis l'étude de la langue naturelle ne passionnait que les linguistes et les grammairiens, aujourd'hui, depuis l'apparition des ordinateurs, l'avènement de l'intelligence artificielle, les capacités de stockage énormes et le potentiel important de traitement des données ont conduit les informaticiens à rejoindre ces chercheurs et faire de la linguistique une des sciences modernes sur le plan social, technologique, économique et industriel. Les mots font bien partie de la langue, mais l'acte de la compréhension de leur signification est une épreuve à laquelle est soumis le destinataire dans un message de langage naturel qu'il doit déchiffrer. Pour aboutir à la compréhension du sens voulu dans les langues entre l'expéditeur et le destinataire, l'individu est obligé de maîtriser les formes de ce vocabulaire, les structures syntaxiques et grammaticales et un tas d'informations et de connaissances pour lier ce vocabulaire et ces structures. En cas d'absence de ces possibilités, l'individu se voit en plein dans l'ambiguïté, et la confusion s'installe quant au choix du sens, du concept ou de la réalité visée par le message linguistique. Cette ambiguïté est due à l'existence de plus d'une possibilité pour interpréter l'unité linguistique. Cette menace de confusion touche tous les niveaux de l'analyse : lexical, grammatical, sémantique, et contextuel. C'est-à-dire que les langages -écrits ou parlés- sont de nature ambigus, sauf que les humains, la plupart du temps se charge de lever cette ambiguïté.

Dans cet article, nous donnons une brève et complète description du phénomène de l'ambiguïté ainsi que les différentes approches de désambiguïsation. Nous proposons ensuite une démarche originale au plan de l'analyse et la modélisation basée sur les techniques de l'aide multicritère à la décision permettant le classement des scénarios de désambiguïsation que nous appliquons en cas de la langue arabe.

II. AMBIGUÏTE : CONCEPT ET UTILITE

La langue naturelle et le moyen de communication le plus redoutable, le plus complexe et le plus précis en raison de sa grande richesse en vocabulaire, de la complexité des structures morphologiques et syntaxiques

¹ TALN : traitement automatique des langues naturelles.

² AMD : Aide Multicritère à la Décision.

et aussi le flux d'informations et de connaissances qui lie le vocabulaire aux différentes structures. L'ambiguïté est le caractère de ce qui possède plusieurs significations. En linguistique et/ou pendant le traitement automatique, l'ambiguïté est un terme qui désigne l'obstacle essentiel dans la compréhension des textes au cours de la lecture et l'analyse.

En langue Arabe la dualité entre le mot et sa voyellation³ suppose un grand accroissement du volume courant de la langue, sachant qu'un mot peut prendre parfois plus d'une vingtaine de formes en fonction de la configuration qui l'accompagne, ce qui entraîne les problèmes les plus complexes dans la compréhension chez l'homme et la machine. On s'intéresse à la multiplicité des formes d'un même vocable de la langue ou à la structure morphosyntaxique et encore au sens de la phrase toute entière ; on appelle ambiguïté, le phénomène qui découle de cette multiplicité.

Cependant on élimine les cas où l'être humain peut surmonter l'ambiguïté, les cas de disponibilité d'un critère linguistique qui permet de désambiguïser et ne sont en aucun cas considérés comme ambigus, voir les exemples ci-dessous.

- Exemple 1: "أكل أيمن الطعام لأنه جائع"
 « akala Aiman eltta3aama li'anahu jaa'i3 »
 "Aïman a pris son mangé parce qu'il était faim"

Explication : La relation de l'adjectif qualificatif⁴ avec le qualificatif⁵ détermine que "جائع" « jaa'i3 » est un adjectif qualificatif épithète de "أيمن" « Aïman », et dans ce cas les linguistes considèrent la phrase non ambiguë.

- Exemple 2: "أكل أيمن الطعام لأنه كان موجوداً"
 « akala Aiman eltta3aama li'anahu kana mawjoudan »
 "Aïman a pris son mangé parce qu'il était disponible"

Explication : Qui est ce qui était disponible Aïman ou le mangé ? Dans ce cas les linguistes considèrent la phrase ambiguë.

A. Pourquoi doit-on lever l'ambiguïté ?

La détermination d'un sens unique et d'une seule catégorie morphosyntaxique pour chaque mot dans le texte traité est nécessaire pour une voyellation correcte de tous les mots du texte, ainsi résoudre la plupart des sujets qui engendrent le traitement automatique de l'arabe. Parmi ces besoins on cite à titre non exhaustif :

- Se permettre une analyse grammaticale correcte.
- Construire des systèmes robustes de questions-réponses dans les langages parlés.

- Bâtir des systèmes de traduction fiables, vu que la plupart des problèmes dans la traduction relèvent de l'ambiguïté.

- Elaborer des systèmes efficaces pour la compression, le résumé et la génération des textes.

B. Le système morphologique de l'arabe : description

Notre étude porte sur l'arabe et en particulier son système morphologique, de ce fait, il nous semble important de donner un aperçu sur les propriétés propres à cette langue. La langue arabe est une langue flexionnelle. Mais la description qui correspond le mieux à son système morphologique est une combinaison de deux descriptions :

a) C'est une langue basée sur l'utilisation de la forme standard pour chaque type de mots. Les noms au sens large ont leurs formes, les verbes aussi. Certaines formes sont communes aux noms et aux verbes. Les changements dans les formes donnent lieu à des changements dans le sens.

b) C'est aussi une langue dérivationnelle. Les dites formes ne seront jamais aptes à accomplir leurs fonctions sans inclure des racines trilitères. Pour distinguer le sens d'un mot dans un contexte donné. Deux facteurs sont liés et ne peuvent être dissociés : *le sens du matériau dont le mot est fait*, et *le sens de la forme qui contient ce matériau*.

C. Désambiguïstation : principe et cas

La langue naturelle est intrinsèquement ambiguë. Cette ambiguïté est source d'explosion combinatoire [8]. Ce phénomène d'ambiguïté est un problème omniprésent dans toutes les langues naturelles. Le code linguistique explique l'existence de plus d'une fonction. Les contextes linguistiques de la langue arabe sont de nature lucide. En dépit de cela, ils nécessitent la présence d'un critère témoin de la fonction du code linguistique dans son contexte. La détermination de ces critères et leur compréhension dépend des capacités de l'individu des points de vue linguistique et du volume de ses connaissances. Plus ces capacités augmentent, plus l'espace de l'ambiguïté diminue aux abords de la compréhension de la langue, et celui de la clarté s'amplifie pour atteindre son point culminant chez le spécialiste. Lorsque l'individu arrive à choisir parmi les diverses solutions celui qui correspond au contexte, on appelle ceci "*désambiguïstation*". La levée de l'ambiguïté ou la désambiguïstation morphosyntaxique a pour objectif la réduction du nombre d'interprétations issues de l'analyse morphologique à l'aide du contexte immédiat⁶.

Il est impossible de décrire l'ensemble des chaînes du discours sans faire appel à des classes de mots. Ainsi, l'article, le nom, le verbe, l'adjectif, ...etc. sont des

³ Soit un ensemble de codes qui assurent un certain nombre de fonctions.

⁴ الصفة.

⁵ الموصوف.

⁶ En fonction du mot qui précède ou qui succède le mot courant.

classes de mots bien connues pour parvenir à une description finie des configurations correctes. Le type d'ambiguïté que nous essayerons de lever dans cette étude est d'une forme morphosyntaxique. Dans la littérature, cette ambiguïté est due généralement à l'attribution de plusieurs informations morphosyntaxiques (verbe/nom, masculin/féminin, singulier /duel / pluriel, ...etc.) à la même unité lexicale

La désambiguïssation sert à lever l'ambiguïté dans le cas où un mot reçoit plus d'une étiquette.

i.e. SI Card(Etiq) = 1 ALORS « Pas de phase de désambiguïssation »

SINON SI Card (Etiq) > 1 ALORS « Phase de désambiguïssation obligatoire »

- m_i est dit ambigu si $\text{Card}(v_i) > 1$, *i.e.* m_i admet plusieurs valeurs grammaticales.
- m_i est dit non ambigu si $\text{Card}(v_i) = 1$, *i.e.* m_i n'admet qu'une seule CG.

III. METHODES DEJA UTILISEES POUR LA LEVEE DES AMBIGUÏTES

Nous n'allons pas faire une liste exhaustive de toutes les tentatives dans le domaine, mais nous nous contentons de citer les travaux les plus connus et les plus récents, en particuliers les travaux de : Bourguignon en 1975, Debili en 1977, Carretero 1979, Al-Nachawati 1981, Kallas 1987, Kupiec et Cutting en 1992 et Chanod 1995, pour résoudre ce problème important. Toutes ces idées généralement se répartissent en deux catégories de modèles, et chaque catégorie englobe une ou plusieurs techniques pour lever l'ambiguïté morphologique. Si plusieurs chercheurs ont préféré l'utilisation des modèles probabilistes pour réaliser l'étiquetage grammatical [2], [5], d'autres ont choisi l'utilisation de règles contextuelles pour atteindre cet objectif.

Dans une comparaison entre l'approche probabiliste et l'approche par contraintes plusieurs chercheurs semblent donner un avantage à la seconde, tout au moins sous la contrainte d'un temps de mise au point limité, une rapidité plus une facilité à mettre en place et aussi une fiabilité au niveau de l'étiquetage [4], [6]. La combinaison de plus d'un modèle est favorable. La plupart des modules de désambiguïssation implémentés dans les systèmes commercialisés combinent entre plusieurs techniques probabilistes et par contraintes pour argumenter la performance de leur étiqueteurs, par exemple, l'étiqueteur MBT, est basé sur la combinaison des deux approches probabilistes et par contraintes.

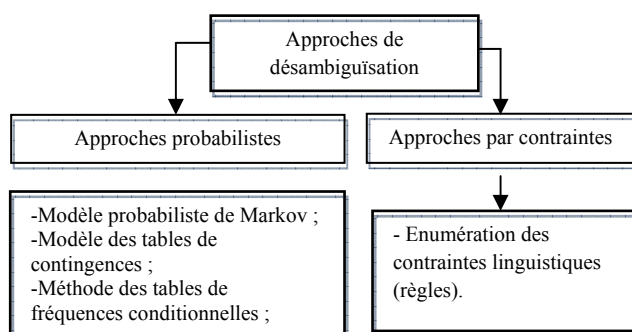


Figure 2: Différentes techniques de désambiguïssation.

ou bien l'inverse, mais il peut se manifester sous différentes formes, comme nous l'illustrons à travers l'exemple :

➤ Exemple : كتب الدرس في القسم .
 Soit : $T = m_1 m_2 m_3 \dots m_i \dots m_n$, *i.e.* $T =$ كتب الدرس في القسم .
 Et soit : $CG = v_1 v_2 v_3 \dots v_i \dots v_n$

- (la écrit) / كُتِبَ (فعل ماضي) / (la leçon) / الدرس (مفعول به) / (à) / في (جار) (la classe) / القسم (مجرور)
- Quismi (nom génétif) Fi (préposition) el-darsa (patient) Kataba (Verbe accompli)
- (à) / كُتِبَ (a été écrit) / مَبْنِي لِلْمَجْهُولِ (la leçon) / الدرس (نائب فاعل) / (à) / في (جار) / القسم (مجرور) / (la classe) / (préposition) el-darsa (pseudo sujet) Kataba (Verbe accompli passive)
- (les livres) / كُتِبَ (مبتدأ) / (de la leçon) / الدرس (مضاف إليه) / (à) / في (جار) / القسم (مجرور) / (la classe) / Quismi (nom génétif) Fi (préposition) el-darsi (annexé) Kutubu (primat)

Figure 1. Exemple d'une ambiguïté Morphologique.

A. Approche par contraintes

Cette approche est basée sur un modèle qui fait intervenir un linguiste, qui va nous permettre d'établir une liste de règles par classe ou par catégorie permettant de lever l'ambiguïté. Ces catégories peuvent être de type : grammaticale, structural, sémantique, logique, ...etc. Les contraintes grammaticales sont surtout utilisées pour la levée de l'ambiguïté due à l'appartenance simultanée de l'unité sémantique à plus d'un modèle grammatical. L'utilisation de la contrainte grammaticale peut suffire à elle seule et parfois l'utilisation de la contrainte sémantique et la contrainte logique s'imposent à ses côtés. Exemple : L'étiqueteur de Brill utilise des règles contextuelles.

A.1. Quelques règles utilisées dans l'analyse de l'arabe

Chaque terme possède des propriétés qui déterminent les mots qui doivent l'accompagner et qui le soutiennent dans l'accomplissement de sa fonction dans la compréhension du texte⁷. Les systèmes à base de contraintes s'intéressent aux règles contextuelles de type :

a) sémantique : comme dans les cas de la relation de l'adjectif avec son qualifié ou la relation de l'annexant⁸ avec l'annexé⁹.

b) grammaticale : généralement traduit un contexte immédiat, par exemple : Une quasi-préposition¹⁰ n'est jamais suivie d'un verbe ou d'une préposition; Une préposition n'est jamais suivie par une autre préposition; Quand deux noms se suivent, le deuxième doit être soit

⁷ C'est l'une des conséquences de la règle statistique.

⁸ مضاف.

⁹ مضاف إليه (Complément de nom).

¹⁰ ظرف (circonstanciel)

un adjectif, soit un annexé ; La préposition "إلى" ou "في" n'est jamais suivie d'un verbe et est toujours suivie d'un nom ; Le mot "الف" n'est jamais suivi d'un nom et est toujours suivi d'un verbe.

B. Approche probabiliste

Dans cette approche le facteur probabiliste et statistique guide les choix, du plus redondant jusqu'au moins redondant. Il s'agit du taux le plus élevé de présence d'une contrainte linguistique, qu'elle soit lexicale, morphologique, syntaxique, morphosyntaxique ou sémantique. Elles sont stockées avec chaque contrainte sous formes de traits fondamentaux. On détermine la contrainte statistique et probabiliste en effectuant une recherche dans des corpus linguistiques pour évaluer le taux de présence de chaque contrainte par rapport à ses semblables. Ce taux est estimé en moyen d'opérations arithmétiques complexes. La levée de l'ambiguïté se fait en utilisant deux types d'informations : la première sur le mot à étiqueter et la seconde contextuelle syntaxique, puis on procède à une combinaison de ces deux informations et à leurs apprentissage¹¹ sur des corpus en général annotés à la main. La technique de Markov constitue un des types de modèles probabilistes les plus utilisés en raison de son efficacité [11] dans l'élaboration d'un module de désambiguïstation. Exemple : l'étiqueteur TnT utilise comme technique le modèle de Markov.

IV. APPROCHE D'ANALYSE MULTICRITERE : PRINCIPE DE LA METHODE PROPOSEE

Notre but est de proposer un nouveau modèle de désambiguïstation qui s'inspire d'une approche mathématique, qui s'appelle AMD (aide multicritère à la décision). Par exploitation multicritère des phénomènes linguistiques, on entend l'utilisation de ces critères et leur prescription dans la compréhension du sens. On commence par faire le choix d'une fonction du code conforme au contexte, pour passer à la prescription des critères linguistiques en délimitant les types de critères utiles dans l'orientation du sens, la collecte des critères à partir de leurs différentes sources, la formalisation des critères sous une forme qui permet à l'ordinateur de les traiter et enfin, l'alimentation de l'ordinateur par ces critères sous une forme adéquate.

Le principe de la méthode de levée des ambiguïtés que nous proposons consiste à réduire, d'emblée, le nombre de scénarios de désambiguïstation en écartant les scénarios dominés (*i.e. scénarios ne possédant aucune*

meilleure évaluation selon tous les critères utilisés) et à classer les scénarios efficaces (*i.e. ceux qui ne sont pas dominés*) afin de faire émerger le meilleur scénario de désambiguïstation, celui qui jouit globalement des scores les plus performants selon les différents critères utilisés.

A. Pourquoi une approche multicritère ?

Le TALN induit souvent des pratiques décisionnelles qui correspondent à un enchaînement de choix. Sachant que le contexte linguistique précis de l'arabe insiste sur la présence des critères qui témoignent de la fonction de plusieurs contraintes (grammaticales, flexionnelles, structurales, sémantiques, logiques et statistiques). Donc l'utilisation des outils d'aide à la décision multicritère s'avère très efficaces. L'avantage d'une telle démarche est de réduire d'emblée le nombre d'étiquettes, en éliminant ceux dominés et de classer les tags efficaces selon un score global calculé. Tous cela est basé sur une bonne définition des critères d'évaluation.

B. Formalisation d'un problème en analyse multicritère

Soit $X = \{x_1, \dots, x_n\}$ l'ensemble des scénarios de désambiguïstation. Ces scénarios sont différents, en nombre fini, et constituent l'intégralité des solutions (étiquettes) possibles.

Pour choisir le meilleur scénario de X , on utilise un ensemble $F = \{f_1, \dots, f_q\}$ qui constitue une famille cohérente¹² de critères. Afin de juger chacun des scénarios de désambiguïstation selon chacun des critères, on définit une fonction d'évaluation de la manière suivante :

$$f_j : X \rightarrow \mathbb{R}, j=1, \dots, q$$

$x \rightarrow f_j(x)$ où $f_j(x)$ représente l'évaluation du scénario x selon le critère f_j .

Chacune de ces fonctions doit être maximisée (ou minimisée) selon le type de critère utilisé.

Scénario idéal : un scénario est dit idéal s'il correspond à la meilleure solution selon tous les critères.

Relation de dominance entre scénarios : on dit qu'un scénario x_1 domine un scénario x_2 si et seulement si : $f_j(x_1) \leq f_j(x_2)$ pour chaque j et avec au moins l'une des inégalités étant stricte.

Scénario efficace : un scénario x dans X est dit efficace si et seulement si aucun autre scénario de X ne le domine [16]. L'ensemble des scénarios efficaces est considéré comme l'ensemble des solutions les plus intéressantes.

Classement des scénarios : L'objectif étant de déterminer le scénario qui jouit globalement des meilleures évaluations. Ainsi, nous calculons pour chaque scénario x_i , un score d'évaluation global $S(x_i)$ qui représente la somme pondérée des différentes évaluations de x_i selon tous les critères.

C. Approche d'analyse multicritère : Principales étapes

¹¹ La technique d'apprentissage et de classification : Un ensemble d'exemples est stocké dans une mémoire ; chaque ensemble contient un mot ou sa représentation lexicale, son contexte (antérieur et postérieur) et la catégorie grammaticale à laquelle il est associé dans chaque contexte. L'analyse se fait de la manière suivante : pour chaque mot de la phrase, le Tagger cherchera un exemple d'emploi analogue dans la mémoire et en déduira sa catégorie grammaticale à laquelle il est associé dans chaque contexte après pondération de l'importance de chaque catégorie.

¹² La famille de critères est Cohérente dans le sens où elle satisfait les trois propriétés suivantes : l'exhaustivité, la cohésion et la non redondance [12].

Les étapes de la méthode proposée sont les suivantes :

- Etape 1 : La mise en place d'un ensemble qui contient toutes les actions ou solutions (dans notre cas il s'agit des étiquettes ambiguës) possibles :

Soit « E » cet ensemble/ $E = \{e_1, e_2, \dots, e_n\}$ où, e_1 : est considéré comme étant une étiquette candidate, qui génère systématiquement une information morphosyntaxique ;

- Etape 2 : Construction d'une famille cohérente de critères : $F = \{f_1, f_2, \dots, f_p\}$;

- Etape 3 : Définir une fonction d'évaluation : générer une fonction d'évaluation pour chaque critère. Le résultat est un tableau d'évaluation appelé matrice d'évaluation.

- Etape 4 : la pondération et l'agrégation des critères :

a) La pondération : consiste à déterminer le poids de chaque critère selon son importance¹³, la méthode de pondération des critères va permettre une discrimination entre les critères en désignant un poids pour chaque critère, ce qui va générer un vecteur de pondération « α ». Pour pondérer les différents critères nous adoptons la méthode de l'Entropie;

b) L'agrégation : Le but est de réduire le nombre d'étiquettes, et de les classer selon leurs scores globaux. Le choix d'une méthode d'agrégation va permettre de normaliser le tableau d'évaluation et facilite une bonne lecture de ce tableau (On obtiendra un tableau « N » ou une matrice « N » normalisé). Afin d'agrèger les différentes évaluations d'un scénario calculées selon les critères retenus, nous proposons la méthode TOPSIS¹⁴.

- Etape 5 : Choisir l'étiquette ayant le plus grand score.

D. Méthode d'agrégation TOPSIS : Fondement

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). TOPSIS est une méthode multicritères développée par Hwang et Yoon en 1981[7]. Il s'agit de réduire le nombre de scénarios de désambiguïsation en écartant les scénarios dominés et de classer les scénarios efficaces selon leurs scores globaux calculés.

D.1. Algorithme

- Etape 1 : Normaliser les performances (i.e. calcul de la matrice de décision normalisée) ;

Les valeurs normalisées « e_{ij} » sont calculées comme suit :

$$e'_{ij} = \frac{g_j(a_i)}{\sqrt{[g_j(a_i)]^2}}$$

¹³ Les critères importants sont ceux qui font la différence (discrimination) entre les solutions; ainsi ces critères auront des poids importants.

¹⁴ TOPSIS : Technique for Order Preference by Similarity to Ideal Solutions

avec $i = 1, \dots, m ; j = 1, \dots, n$. où $g_j(a_i)$

correspondent aux valeurs déterministes des actions i pour le critère j .

- Etape 2 : Calcul de la matrice de décision normalisée pondérée (i.e. Calculer le produit des performances normalisées par les coefficients d'importance relative des attributs). Les éléments de la matrice sont calculés comme suit :

$$e''_{ij} = \pi_j e'_{ij} \quad \text{Avec } i = 1, \dots, m ; j = 1, \dots, n. \\ \text{et } \pi_j \text{ est le poids du } j^{\text{ème}} \text{ critère}$$

$$\sum_{j=1}^n \pi_j = 1$$

- Etape 3 : Détermination des solutions (profils) idéale (a^*) et des solutions anti-idéale (a_*) ;

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

$$a_* = \{ e^*_{j^*}, j = 1, \dots, n \} = \{ e^*_1, e^*_2, \dots, e^*_n \} ;$$

$$e^*_{j^*} = \text{Max}_i \{ e''_{ij} \}$$

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

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

$$e_{j^*} = \text{Min}_i \{ e''_{ij} \}$$

- Etape 4 : Calcul des mesures d'éloignement (i.e. Calculer la distance euclidienne par rapport aux profils a^* et a_*) ; L'éloignement entre les alternatives est mesuré par une distance euclidienne de dimension n . L'éloignement de l'alternative i par rapport à la solution idéale (a^*) qui peut être assimilé à la mesure d'exposition aux risques et donné par :

$$D^*_i = \sqrt{\sum_{j=1}^n (e''_{ij} - e^*_j)^2} \quad i = 1, 2, \dots, m$$

$$D_{i^*} = \sqrt{\sum_{j=1}^n (e''_{ij} - e_{j^*})^2} \quad i = 1, 2, \dots, m$$

- Etape 5 : Calculer un coefficient de mesure du rapprochement au profil idéal :

$$C_i^* = \frac{D_{i^*}}{D^*_i + D_{i^*}} \quad i = 1, \dots, m \\ \text{avec, } D^*_i + D_{i^*} \leq C_i^* \leq 1$$

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

E. Méthode de pondération Entropie

La méthode Entropie 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 (dans notre cas se sont les étiquettes).

E.1. Algorithmie

L'entropie d'un critère « j » est calculée par la formule[12] :

$$E_j = -K \sum_{i=1}^n X_{ij} \log(X_{ij})$$

où K est une constante choisie de telle sorte que, pour tous « j », on a $0 \leq E_j \leq 1$,

par e.g. : $K = 1 / \log(n)$ (n étant le nombre de scénarios de désambiguïsation).

L'entropie E_j est d'autant plus grande que les valeurs de « e_j » sont proches. Ainsi, les poids seront calculés en fonction de la mesure de dispersion (opposée de l'entropie) :

$$D_j = 1 - E_j$$

Les poids seront ensuite normalisés par :

$$W_j = D_j / \sum_j D_j$$

V. PRESENTATION DE LA SOLUTION ET APPLICATION

Notre solution suit la démarche suivante :

-Etape 1 : Construction de la liste des étiquettes. Cette liste est construite directement après une analyse morphologique ambiguë (plusieurs solutions possibles) ce qui va générer l'ensemble E.

Ainsi pour la phrase : "ذهب محمد إلى المدرسة" "Mohamed est parti à l'école".

Après la reconnaissance (une analyse morphosyntaxique), les schèmes générant ce mot peuvent être différents, i.e. nous sommes devant un cas de mot ambigu "ذهب" qui peut être un nom commun ذَهَبٌ de schème (فَعَلٌ) ou un verbe (Vtype1 فَعَلَ , Vtype2 فَعِلَ , Vtype3 فَعُلَ , Vtype4 فَعَلْ , Vtype5 فَعِلْ), dans ce cas l'ensemble E sera :

$E = \{\text{Non commun, Vtype1, Vtype2, Vtype3, Vtype4, Vtype5}\}$

-Etape 2 : Afin de construire une famille cohérente de critères F, nous proposons trois critères de base pour discriminer entre les scénarios d'étiquetage : le critère de

concordance de voyelles à l'intérieur du mot, critère de fréquence et le critère contexte structural.

- Critère de concordance des voyelles

Ce critère va utiliser la position des voyelles pour lever l'ambiguïté, il s'agit d'un critère à maximiser. La fonction d'évaluation qui va avec c'est l'addition (+), de telle manière, qu'une bonne position d'une voyelle vaut un (1), après application du critère on aura :

Non commun (1+1+1=3), Vtype1(1+1+1=3), Vtype2(1+0+1=2), Vtype3(1+0+1=2), Vtype4(0+0+1=1), Vtype5(1+0+0=0).

- Critère de fréquence

Ce critère favorise le scénario dont les caractéristiques ont la plus grande fréquence. Le score d'un scénario x selon ce critère représente son taux d'apparence calculé sur la base de l'étude statistique dans le corpus utilisé donc il s'agit d'un critère à maximiser. On distingue les six cas possible auxquels on assigne les scores spécifiques : Non commun (0,75), Vtype1 (0), Vtype2(0), Vtype3(0), Vtype4(0,25), Vtype5(0,5).

- Critère paramètre structural

Supposant que le mot "ذهب", il s'agit d'un nom commun ça équivaut en français à un métal précieux "l'or" et comme dans le cas présent suivi d'un nom propre "محمد", il est alors un syntagme nominal composée d'un *annexant* et d'un *annexé*. Le paramètre structural (règle grammaticale) n'admet pas que ce composé soit suivi d'une préposition et un nom génitif¹⁵ (à finale réduite) par "إلى المدرسة" "à l'école". Donc notre supposition initiale est fautive ce qui rend la phrase "ذهب محمد إلى المدرسة" "l'or de Mohamed à l'école" n'a pas de sens point de vue structure. Cette conclusion va nous imposer le score zéro pour l'étiquette Non commun (0) au mot "ذهب". Ce critère donne le score (1) aux étiquettes de verbe (i.e. Vtype1 (1), Vtype2(1), Vtype3(1), Vtype4(1), Vtype5(1)).

Prenons toujours le mot "ذهب" dans la phrase : ذهب "L'or de Mohamed est éblouissant". Dans cet exemple, le contexte présume qu'il s'agit d'un verbe. Mais la règle structurale réfutera cette idée. À ce titre, le contexte exige que le mot "ذهب" soit un nom, parce que le syntagme nominale est suivie de "براق" "éblouissant" qui est une description du nom "ذهب" du point de vue grammaticale c'est-à-dire un attribut¹⁶. Les contraintes sémantiques admettent que "براق" soit un adjectif qualificatif attribut du nom "ذهب". Et si la description se rapportait à "محمد", la règle structurale admet que "ذهب"

¹⁵ جار ومجرور

¹⁶ خبر Khabar

soit un *verbe* comme dans la phrase "ذهب محمد ضاحكا" qui veut dire « **Mohamed est parti en riant** ». Ici la description joue le rôle d'un *accusatif d'état*¹⁷. Il s'agit dans ce cas d'un *verbe*, un *sujet* et d'un *accusatif d'état*.

- Etape 3 : Utilisation des méthodes de pondération et d'agrégation sur le tableau d'évaluation. Puis on va faire le produit entre le tableau normalisé et le vecteur des poids, se qui va générer un tableau V normalisé et pondéré. $V = \alpha.R$

- Etape 4 : Classification des étiquettes. Le classement des étiquettes est effectué selon un ordre décroissant de leurs scores globaux. Le calcul d'un score global, se fait en utilisant la formule :

$$S(e_{ij}) = \sum_j V_{ij} \quad i = 1, \dots, n; j = 1, \dots, q$$

$$V = \alpha.R$$

VI. CONCLUSION

L'étiquetage morphosyntaxique est considéré aujourd'hui, comme étant une partie vitale dans n'importe quelle application de TALN (traducteur automatique, correcteur orthographique, système de résumé automatique...). Ainsi la performance d'une application dépend directement de la performance de l'étiqueteur, et pour avoir une fiabilité irréprochable, on doit s'intéresser à trois points, qui sont : une bonne phase de segmentation, une bonne organisation des unités lexicales (base de données ou bien dictionnaire) et un module de désambiguïsation très fiable. C'est ce dernier point qu'est le centre d'intérêt de notre étude. Nous avons essayé à travers cette étude de décrire l'ambiguïté morphologique et de mentionner les différentes approches existantes telles que l'utilisation des modèles probabilistes ou l'introduction des modèles par contraintes contextuelles.

Mais le but principal de cet article reste la présentation d'une nouvelle approche mathématique fondée sur l'AMD, permettant le classement multicritère des scénarios de désambiguïsation en vue d'en faire émerger le meilleur. Cette méthode a l'avantage de réduire les scénarios dominés et de classer le reste selon différents critères d'évaluation. Cette technique bien qu'elle est peu exploitée, montre que la voie d'une analyse multicritère dans le TALN est intéressante. Ce type de systèmes, offre une alternative aux systèmes basés sur une approche probabiliste et peut être un complément indispensable au modèle par contrainte contextuelle.

REFERENCES

[1] Brault F., Forces et faiblesses de l'utilisation de trigrams dans l'étiquetage automatique du français. Exploration à partir des homographes de type verbe-substantif, Maîtrise en linguistique, Univ. Laval, 2007.

- [2] Cutting D., Kupiec J., Pedersen J. and Sibun P., A practical part-of-speech tagger, in proceedings of the third conference on applied natural language processing, 1992.
- [3] Daelemans, Walter et al. MBT: Memory-Based Tagger, version 1.0, Reference Guide. ILK Technical Report –ILK 02-09. October, 29, 2002.
- [4] D. Allotti, C. Ponsard, « Exposé sur l'étiqueteurs Statistiques et étiqueteurs par contraintes », 2005.
- [5] El-Bèze M, Merialdo B., Roseron B. et Derouault A. M., Accentuation automatique de textes par des méthodes probabilistes, Techniques et Sciences informatique, pp. 797-815, 1994.
- [6] J.P. Chanod, P. Tapanainen, « Les étiqueteurs statistiques et les étiqueteurs par contraintes », 1995.
- [7] Hwang C. R, Yoon K., Lecture Notes in Economics and Mathematical Systems, Springer-Verlag Berlin Heidelberg, New York, 1981.
- [8] Lallich-Boidin G., Henneron G. et Palermi R., Analyse du Français : Achèvement et implantation de l'analyseur morphosyntaxique, Les cahiers du CRISS, N° 16, 1990.
- [9] Lecomte, Josette. Le Catégoriseur Brill 14-JL5/WinBrill-0.3. InaLF/CNRS, décembre 1998.
- [10] Med. Benhamou, Y. Hoceini, المعالجة الآلية للبيانات اللغوية الطبيعية : اللغة العربية نموذجاً, JeTIC2006, CUB.
- [11] Merialdo B., Tagging english text with a probabilistic model, Computational linguistics, 1994.
- [12] Pomerol J.-C., Romero S. B., Choix multicritère dans l'entreprise : principes et pratique, Hermès, 1993.
- [13] M. Constant, « Etiquetage morphosyntaxique probabiliste », cours pour Master en informatique, Université Paris-Est Marne-la-Vallée, 2007.
- [14] Roy B., Méthodologie Multicritère d'Aide à la Décision, Collection Gestion, Ed. Economica, 1985, Paris.
- [15] Santorini, Beatrice. Part-of-Speech Tagging Guidelines for the Penn Treebank Project. March 15, 1991. [téléchargeable sur le site de TreeTagger]
- [16] Vincke P., L'aide multicritère à la décision, SMA, Université de Bruxelles, Bruxelles, 1989.
- [17] Y. Hoceini, N. Manouni, L'informatique et la langue arabe et les solutions importées, Vol N°1, WATA : World Association of Arab Translators and Linguists, . Janvier 2007. <http://arabswata.info/mag/ab7ath/p3.htm>
- [18] Y. Hoceini, Mohamed Benhamou, Initiative en TAAA¹⁸: présentation d'une expérience de recherche de GTALA-CUBA, revue N°2, Hawliyat CUB¹⁹, 2007.
- [19] Y. Hoceini, Procédés de reconnaissance de formes en analyse morpholexicale de l'arabe, JeTIC2006, CUB.

¹⁷ حال

¹⁸ Traitement Automatique de l'Ambiguïté de l'Arabe (T3A)

¹⁹ Centre Universitaire de Bechar