TALN '99
6ème Conférence sur le Traitement Automatique des Langues Naturelles
Corse, 12-17 juillet 1999

Algorithmes pour la correction des erreurs orthographiques en arabe

Chiraz Ben Othmane Zribi
Université Paris-sud, Orsay, France
chiraz@apexmail.com

Adnane Zribi
Université Tunis III, I.S.G., Tunisie
Adnane.Zribi@isg.rnu.tn

Résumé

Nous traitons dans ce papier du problème de la détection et de la correction des graphies fautives dans les textes arabes. Nous commençons par présenter une expérience visant à mesurer de manière comparative la difficulté du problème pour l'arabe, le français et l'anglais. L'idée est d'évaluer le degré de "ressemblance" (proximité) des mots au sein de chaque langue. Ensuite les algorithmes de base de notre méthode de correction sont présentés.

Introduction

La détection-correction des erreurs orthographiques est un problème qui a donné lieu à de nombreuses études. Pour des langues telles que le français ou l'anglais, des techniques de correction automatique des erreurs existent et donnent des résultats que l'on juge généralement satisfaisants. Pour le cas de l'arabe, l'application de ces techniques pose problème. En effet, les particularités de la forme textuelle arabe: agglutination des enclinomènes (enclitiques et proclitiques) aux formes d'une part, et voyellation d'autre part, font de l'arabe un cas qu'il faut étudier à part. L'on pourrait penser, notamment à cause de l'existence de correcteurs orthographiques commercialisés que le problème est résolu. En réalité, les solutions proposées sont loin d'être satisfaisantes et de réelles difficultés sont encore posées. Les correcteurs disponibles présentent de sérieuses défaillances. En particulier, les erreurs mettant en cause les voyelles ne sont pas corrigées. Pire, elles ne sont même pas détectées. D'autres insufffisances (bruit et silence) sont fréquemment observées aussi bien au niveau de la détection des erreurs qu'à celui de leur correction [Ben Othmane Zribi, 1998].

Difficultés

Selon la langue, quand on veut corriger automatiquement les graphies fautives, l'on ne se trouve pas confrontés aux mêms difficultés. Pour l'arabe, en plus de la voyellation et de l'agglutination, nous avons trouvé, grâce à l'expérience que nous présentons ici, que *"les mots arabes sont très voisins du point de vue lexical"*.

Cette expérience consiste à provoquer automatiquement pour chaque mot de la langue toutes les opérations d'édition (substitution d'une lettre par une autre, ajout d'une lettre, suppression d'une lettre et interversion de deux lettres adjacentes) qu'il est susceptible de subir. Un ensemble de formes est ainsi généré. Il s'agit d'y dénomber les formes correctes. Nous déterminons ainsi ce que nous avons appelé *le nombre de mots approchants ou lexicalement voisins*. Une moyenne calculée sur tous les mots d'une langue (d'un dictionnaire) nous donne une idée de la proximité des mots dans cette langue. Comme le montre le tableau ci-après, le

nombre moyen de formes approchantes pour l'anglais est de ***3*** et pour le français ***3.5***. Pour l'arabe non voyellé, ce nombre est de ***26.5***. Les mots arabes seraient donc beaucoup plus proches les uns des autres que les mots français et anglais.

	Anglais			Français			Arabe		
	Min.	Moy.	Max.	Min.	Moy.	Max.	Min.	Moy.	Max.
Générés automatiquement	82	505	1483	108	892	1881	106	458	1187
dont Reconnus	0	3	54	0	3.5	45	0	26.5	185
Proportion moyenne	**0.59%**			**0.39%**			**5.79%**		

Mots lexicalement voisins, tableau comparatif [1]

Ces comptages nous informent également de la probabilité de tomber sur un mot correct quand on commet une erreur sur un mot donné. C'est le cas lorsque, par exemple, au lieu de taper au clavier le mot " كتـــب ", l'on tape le mot " كثب" ou que, en R.O.C., l'on reconnaît la forme " ســـاء " à la place de " شاء ". Ainsi, on voit que cette probabilité (ce risque) pour un mot arabe est ***10*** fois plus grande que pour un mot anglais et ***14*** fois plus grande que pour un mot français. Il est à signaler toutefois que ces résultats sont "en définition" et non "en usage", c'est à dire qu'ils résultent de comptages effectués sur des dictionnaires et non sur des données textuelles. Ils n'en sont pas moins, pensons-nous, édifiants.

Cette proximité des mots arabes a une double conséquence. D'abord à la détection où les mots reconnus corrects peuvent facilement receler une erreur. Ensuite à la correction où le nombre de candidats pour une forme erronée risque d'être démesuré. A priori, l'on peut penser que nous aurons en moyenne ***27*** formes candidates à la correction de chaque erreur et que ce nombre peut atteindre un maximum de ***185***. Ce serait déjà considérable. Mais il ne faut pas oublier que le phénomène d'agglutination des enclinomènes aux formes augmentera encore ces valeurs. En réalité, ces chiffres risquent, dans certains contextes, de remettre en cause l'intérêt même d'un correcteur orthographique en arabe. Dans un logiciel de traitement de textes, par exemple, entre choisir le mot correct dans une liste de 27 candidats et corriger simplement au clavier la forme erronée, il n'est pas évident que le premier choix l'emportera toujours.

Méthode

Notre méthode de vérification et de correction des mots arabes se base sur l'utilisation d'un dictionnaire. Ce dernier est consulté aussi bien pour décider de l'appartenance des mots à vérifier au vocabulaire que pour générer les mots candidats à la correction des mots erronés.

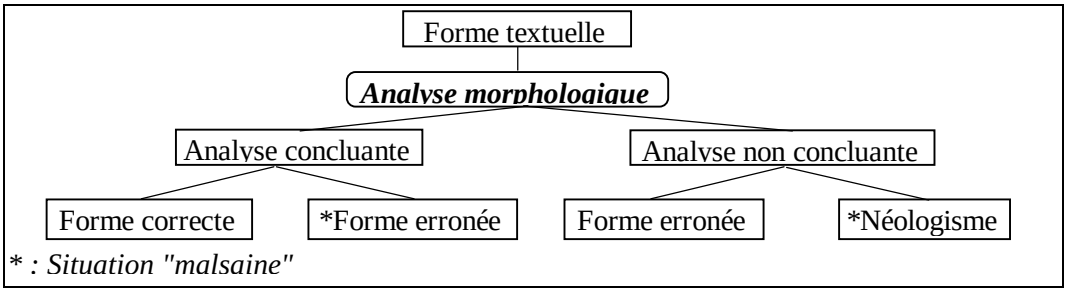
Le dictionnaire que nous utilisons est un dictionnaire de formes fléchies voyellées (1 600 000 entrées) orienté vers les applications classiques du traitement automatique du langage naturel (analyse morpho-syntaxique, indexation automatique, …). On trouve dans ce dictionnaire des mots tels que: ("kataba": a écrit) كَتَـــبَ , ("yaktoubouna": écrivent) يَكْتُبُونَ , ("madrasatâni": deux écoles) مَدْرَسَتَانِ, ("madârison": écoles) مَدَارِسُ, … accompagnés de diverses informations linguistiques les décrivant. À cause de l'agglutination des proclitiques (articles, prépositions, conjonctions) et des enclitiques (pronoms) aux formes fléchies, ce dictionnaire ne suffit pas pour reconnaître les mots telsqu'ils se présentent dans les textes arabes (ex: "va

[1] Cette expérience a été menée sur *84 570* formes anglaises, *293 573* formes françaises et *577 546* formes arabes non voyellées.

l'écrire": سَيَكْتُبُهُ, "et son crayon": بِقَلَمِهِ, etc.). Les ambiguités de découpages résultant de ces situations rendent difficile la reconnaissance des formes fléchies et des enclinomènes. Nous avons donc accompagné ce dictionnaire d'algorithmes permettant l'analyse morphologique des formes textuelles. Cet analyseur morphologique [Zouari 89] utilise, en plus du dictionnaire des formes fléchies, un petit dictionnaire qui contient tous les enclinomènes (90 entrées) et applique un ensemble de règles pour rechercher tous les découpages possibles en proclitique, radical et enclitique.

Détection des erreurs

Au niveau d'analyse où nous nous situons, le seul moyen dont nous disposons pour détecter si une forme textuelle comporte une erreur ou pas est d'effectuer son analyse morphologique. Si l'analyse n'aboutit pas, nous sommes alors dans l'une des deux situations suivantes. Soit la forme textuelle est effectivement erronée, soit il s'agit d'un néologisme (pour l'analyseur). Il nous est cependant impossible de faire la distinction entre les néologismes et les véritables erreurs. Si l'analyse morphologique réussit, nous sommes encore dans l'une ou l'autre de deux possibilités. Ou la forme est correcte, ou elle comporte malgré tout une erreur qui a donné lieu à une forme textuelle acceptable par l'analyseur morphologique. Il est également impossible, à moins de passer à d'autres niveaux d'analyse mettant en œuvre d'autres types de connaissances, de discerner les formes correctes des formes comportant des erreurs mais acceptées par l'analyseur. Le schéma suivant résume les situations que l'on peut rencontrer.



Détection des erreurs

Correction des erreurs

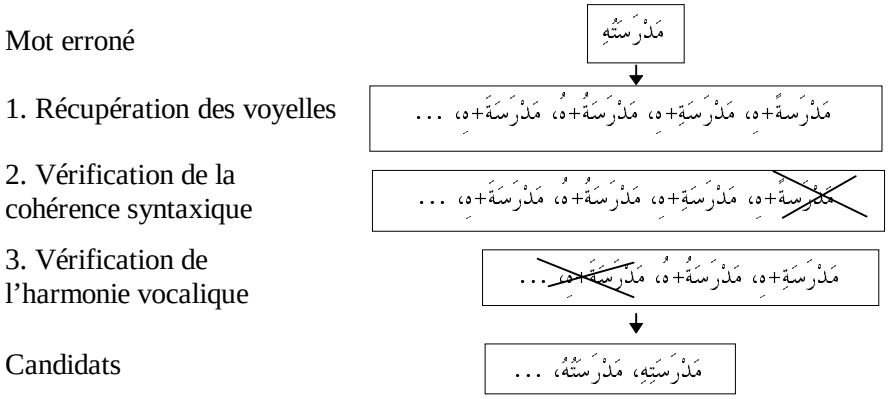
Erreurs de voyellation

Il y a erreur de voyellation quand le détecteur d'erreurs (qui est ici l'analyseur morphologique) relève l'une des deux situations suivantes:

- Une dissemblance entre les voyelles de la forme textuelle avec celles du dictionnaire. Les voyelles incriminées peuvent être celles de la forme fléchie ou celles des enclinomènes qui s'y agglutinent. *Ex:* مدرسة، مدرسة، ومدرسة، مدرسته
- Une voyellation incohérente de la forme textuelle. Pris séparément, le radical et les enclinomènes qui viennent s'y agglutiner sont corrects. Pourtant, le proclitique ou l'enclitique reconnu ne peut être collé à la forme fléchie étant donnée la voyellation qu'a cette dernière. *Ex:* بالمدرسة، ومدرسته، المدرسة، بضرب

Qu'elles relèvent de l'une ou de l'autre de ces deux situations, les erreurs de voyellation sont les erreurs les plus faciles à corriger. En effet, l'organisation que nous avons adoptée pour le dictionnaire des formes fléchies et pour le dictionnaire des enclinomènes s'y prête particulièrement bien. Dans ces dictionnaires, les entrées sont respectivement les formes fléchies non voyellées et les enclinomènes non voyellés. À chaque entrée correspondent, entre

autres informations, toutes les voyellations qu'elle peut avoir. Ainsi la forme candidate à la correction d'une forme textuelle erronée dont l'erreur porte sur les voyelles est cette même forme textuelle sauf qu'elle est revoyellée en considérant les voyellations de ses composantes. Ces voyellations sont d'abord recueillis dans les dictionnaires, les combinaisons non valides sont éliminés par la suite. Cette validation s'effectue en considérant la compatibilité syntaxique, et l'harmonie vocalique (la compatibilité de la voyellation de l'enclinomène avec celle qui la précède). Le schéma suivant résume la démarche suivie:



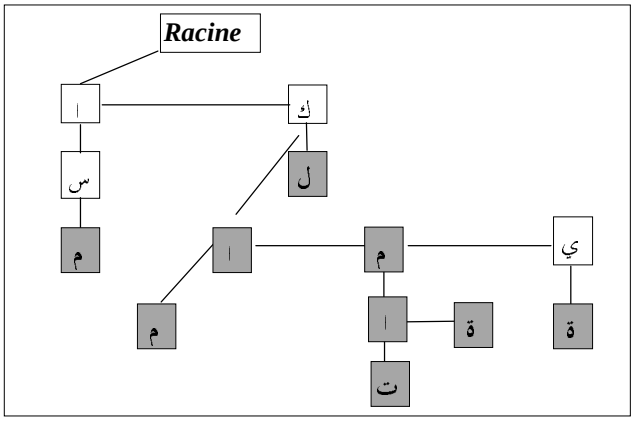
Correction des erreurs de voyellations

Erreurs de consonnes

Notre méthode de correction dépend fortement de la structure de nos dictionnaires (le dictionnaire des formes fléchies et le dictionnaire des enclinomènes). Aussi, nous donnons dans ce paragraphe une description sommaire de cette structure.

Structure du dictionnaire

La structure de notre dictionnaire de formes fléchies est arborescente (idem pour le dictionnaire des enclinomènes). Voici comment sont disposés les mots suivants: اسم، كل، كلا، كلام، كلم، كلما، كلمة، كلمات، كلية.

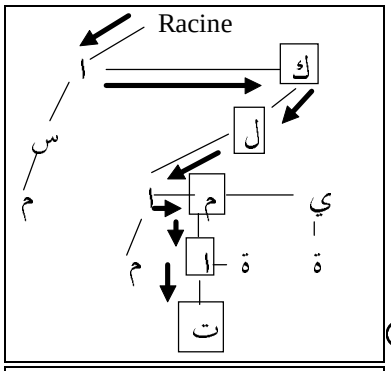


Structure arborescente.

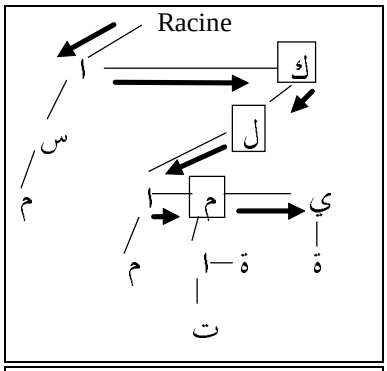
Les mots dans l'arbre sont triés. Chaque noeud de l'arbre correspond à une lettre dans un mot. Les nœuds sont reliés entre eux par deux liens: le lien fils et le lien frère. Certains nœuds (les nœuds en fond grisé dans le schéma ci-dessus) contiennent des lettres de fin de mots.

Le principal atout de cette structure arborescente est qu'elle permet avec grande facilité toutes sortes d'accès. Cela va du simple accès exact et direct à une entrée jusqu'à l'accès aux entrées qui y sont incluses ou qui l'englobent. Pour rechercher un mot dans l'arboresence, on procède à une recherche en profondeur d'abord. La lettre courante est recherchée d'abord dans

le fils du nœud courant, ensuite dans ses frères quand elle n'est pas trouvée. Si la recherche d'un mot aboutit, on obtient *un chemin de nœuds complet*, contenant toutes les lettres du mot recherché, si par contre la recherche échoue on obtient un *chemin partiel* composé d'une partie plus au moins grande du mot recherché. Nous donnons sous forme de schéma la procédure de recherche. Deux cas sont prévus. Dans le premier cas, le mot recherché est trouvé. Dans le second, il ne l'est pas.



Mot recherché:كلمات
Résultat: *Le mot est trouvé*



Mot recherché:كلهات
Résultat*: Le mot n'existe pas*

Correction d'une forme isolée non voyellée

Avant de décrire l'algorithme qui détermine les mots candidats à la correction, d'un mot erroné, deux points sont à préciser:

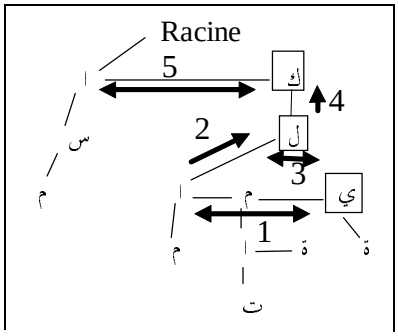
- Nous supposons que dans un mot erroné, il n'y a qu'une seule erreur. Cette erreur est provoquée par l'une des quatre opérations d'édition (substitution, insertion, suppression et interversion). Nous ne savons bien entendu pas, a priori, laquelle de ces opérations d'édition est celle qui a provoqué l'erreur.
- Quand l'algorithme de recherche échoue et qu'il rend un chemin partiel tel que précisé plus haut, l'erreur recherchée se situe forcément dans la partie du mot couverte par ce chemin.

L'algorithme que nous décrivons dans ce paragraphe est un algorithme d'accès particulier au dictionnaire. Il accepte en entrée un mot (le mot recherché) et un chemin partiel (issu de sa recherche infructueuse). Son résultat est une liste formée de l'ensemble des mots voisins (ne différant que par une seule opération d'édition) au mot en entrée et qui sont en fait les candidats à sa correction. Le principe général utilisé est simple.

On parcourt le chemin partiel et on cherche les mots voisins, en supposant successivement que dans la forme erronée il y a eu une opération de substitution d'une lettre par une autre, une opération d'ajout d'une lettre, une opération de suppression d'une lettre et une opération d'interversion de deux lettres consécutives. L'algorithme se présente donc en quatre modules correspondant aux quatre opérations d'édition pouvant être à l'origine de l'erreur dans le mot. Chaque module fournit une liste de candidats. Ces listes seront assemblées pour donner la liste définitive des candidats à la correction du mot erroné.

Nous présentons à titre d'exemple, le module permettant de construire les candidats à la réparation d'une erreur de substitution.

Un exemple: Correction d'une erreur de substitution



Le but de cette partie de l'algorithme de correction est de trouver la lettre qui s'est substituée à une autre et de chercher toutes les lettres qu'elle a potentiellement remplacées. En trouvant ces dernières, l'on trouve de fait les mots candidats à la correction de cette erreur.

Nous savons que la lettre substituante appartient au chemin partiel qui est ici une donnée. Nous savons par ailleurs que les lettres potentiellement remplacées sont à trouver dans l'arborescence. Pour les retrouver, il suffit de parcourir un à un les nœuds appartenant au chemin partiel et le remplacer par tous ses frères (les nœuds qui dans l'arborescence appartiennent au même niveau et qui ont le même père). À chaque remplacement de nœud, il faut vérifier que la partie restante du mot est trouvée.

Correction d'une forme textuelle non voyellée

L'algorithme que nous venons de décrire permet de trouver les mots candidats à la correction d'une forme isolée erronée. Il pourrait être appelé pour corriger également des formes agglutinées erronées, mais il faudrait pour cela que le dictionnaire utilisé contienne l'ensemble des formes agglutinées arabes. Comme le dictionnaire que nous utilisons ne contient que les formes fléchies, nous avons adapté notre algorithme pour qu'il puisse prendre en compte le phénomène de l'agglutination des enclinomènes. En réalité, nous avons construit une nouvelle version de notre analyseur morphologique (dite "version tolérante") qui détecte les erreurs dans les formes textuelles mais reste insensible à ces erreurs.

Conclusion

Deux principaux points ont été abordés. Le premier a consisté à décrire une expérience, que nous pensons fort instructive, permettant de juger (de manière comparative avec deux autres langues) la proximité ou la ressemblance des mots de l'arabe. Nous avons pu mettre en évidence que les mots arabes sont beaucoup plus proches les uns des autres que les mots du français et de l'anglais. Cette proximité a des répercussions au niveau de la fréquence des erreurs orthographiques mais aussi et surtout au niveau des méthodes de leur détection et leur correction automatiques. Le deuxième point traité a été la présentation de nos algorithmes de base pour la correction des erreurs orthographiques au sein des mots arabes.

Références

[Ben Hamadou, 93] Ben Hamadou A. "Vérification et correction automatiques par analyse affixale des textes écrits en langage naturel: le cas de l'arabe non voyellé", Thèse d'état de l'univeristé de Tunis I, soutenue le 12 mars 1993.

[Ben Othmane Zribi, 98] Ben Othmane Zribi C. "De la synthèse lexicographique à la détection et à la correction des erreurs arabes ", Thèse de doctorat, Université de Paris XI, Orsay, 1998.

[Zouari, 89] Zouari L. "Construction d'un dictionnaire orienté vers l'analyse morpho-syntaxique de l'arabe, écrit voyellé ou non voyellé", Thèse de doctorat, Université de Paris XI, Orsay, 1989.