Une méthode universelle de restauration des signes diacritiques

Victoria Bobicev, Victoria Lazu, Liviu Carcea
Technical University of Moldova
victoria.bobicev@rol.md, lazu_vic@yahoo.com, carcea@mail.utm.md

***Résumé* — Cet article présente une méthode universelle de restauration des signes diacritiques. Presque toutes les langues européennes utilisent les signes diacritiques et leur absence dans le texte est un problème commun. La méthode proposée est indépendante du langage. Elle applique un modèle statistique du texte au niveau des lettres/caractères et n'a besoin que d'un corpus d'un volume réduit pour obtenir de bons résultats. La méthode peut être facilement adaptée aux autres langues; nous avons évalué la méthode pour quatre langues: le roumain, le français, l'italien et l'espagnol. Pour toutes les langues la qualité de la restauration des signes diacritiques est supérieure à 99 pour cent au niveau des lettres. L'algorithme est implémenté en Perl, et peut être utilisé dans tout système de traitement de texte. Nos plans d'avenir comprennent l'évaluation de la méthode pour les autres langues européennes, notamment pour les langues est-européennes dans lesquelles l'utilisation des signes diacritiques est plus intense.**

***Mots-clés* – correction du texte, les signes diacritiques, la méthode d'apprentissage automatique, PPM.**

I. INTRODUCTION

Le problème des signes diacritiques existe pour la majorité des langues européennes. Dans [7] sont données les langues qui sont confrontées au problème des signes diacritiques dans les textes.

Dans de nombreux textes présentés sur Internet, les lettres avec des signes diacritiques sont remplacées par leurs variantes sans aucun signe, parce que le changement de la codification peut conduire souvent à leur détérioration. De tels textes sont désagréables à lire pour l'humain, mais celui-ci est quand même en mesure d'en comprendre malgré l'absence ou la substitution des signes diacritiques. En revanche, le problème est beaucoup plus grave pour la machine: si une personne peut rétablir le texte grâce à son sens, ce n'est pas le cas de l'ordinateur, et le problème consistant à corriger un tel texte n'est pas trivial.

II. LES TRAVAUX ANTÉRIEURS

Le problème de la restauration des signes diacritiques a été abordé par de nombreux chercheurs. Comme il est mentionné dans [7], les méthodes basées sur des dictionnaires ont une précision qui dépasse 90% [6], [15]. Mais, premièrement, pas tous les mots font partie des dictionnaires et, deuxièmement, pour beaucoup de mots plus endommagés se trouvent quelques variantes de mots corrects dans le vocabulaire. Pour sélectionner la version appropriée ont été proposée diverses méthodes statistiques.

Les méthodes développées jusqu'à présent pour résoudre ce problème sont généralement basées sur des dictionnaires et divers processeurs lexicaux et/ou syntaxiques. Par exemple, dans [16] a été présentée une méthode basée sur un dictionnaire, ce qui conduit à la restauration de près de 90% des signes diacritiques dans les textes français et espagnols. Dans [6] pour rétablir les signes diacritiques a été utilisé un annotateur morphologique, qui avait un gros dictionnaire. Dans [10] a été proposée une méthode basée sur des mots voisins. Dans [5] on a utilisé une chaine de Markov sur la base de séquence de trois mots et le résultat a été d'environ 99% pour les textes français. Une méthode basée sur la séquence de mots voisins a été présentée dans [9]. Dans [16] plusieurs méthodes ont été présentées, dont la plupart s'appuient sur des dictionnaires et le choix du candidat approprié parmi les mots voisins.

Pour la langue roumaine ont été proposées deux méthodes pour restaurer les signes diacritiques. La première a été décrite par [13], et est basée sur un analyseur morphologique. Le choix du mot juste est effectué sur la base de son contexte et des caractéristiques morphologiques en utilisant la chaine de Markov. Une méthode originale a été décrite dans [7]. La méthode permet de résoudre le problème, non au niveau du mot mais au niveau des lettres. Il s'agit d'une méthode d'apprentissage automatique, mise en œuvre en utilisant TiMBL [3]. La précision obtenue est de 98-99% de signes diacritiques correctement restaurés au niveau des lettres.

La méthode de restauration des signes diacritiques utilisant TiMBL a été présentée aussi dans [14], pour la langue africaine gĩkũyũ. Les auteurs ont également effectué des expériences pour l'allemand, le français et le néerlandais. Les pourcentages de précision des mots sont: pour le néerlandais c'est une très grande précision de 99.8% en raison de l'utilisation limitée des signes diacritiques dans cette langue, puis vient l'allemand avec 95.3%, le français – 89.3% et pour gĩkũyũ 91.4%. L'approche d'apprentissage automatique présentée dans [4] est une recherche plus approfondie sur un éventail plus large de langues africaines (cilubà, gĩkũyũ, kĩkamba, maa, sesotho sa leboa, tshivenda et yoruba) et des expériences sur les langues européennes (le tchèque, le néerlandais, le français, l'allemand et le roumain), ainsi que le pinyin, le vietnamien et le chinois.

Dans cet article est présentée une autre méthode pour restaurer les signes diacritiques au niveau des lettres basée sur le modèle PPM (prediction by partial matching -

prédiction par correspondance partielle) de compression statistique des textes.

III. PPM

Les chercheurs ont observé que les modèles conçus pour compresser le texte sont très similaires avec les modèles développés pour son traitement. La compression réduit la quantité de mémoire requise pour la tenue de dossiers dans l'ordinateur et le temps requis pour la transmission de données. Les meilleures méthodes sont basées sur un codage statistique [2]. Dans la compression statistique à chaque symbole est attribué un code sur la probabilité du texte. Les symboles avec une grande probabilité obtiennent des codes courts, les symboles avec une petite probabilité ont des codes longs.

Le modèle adaptable de compression de texte PPM (prediction by partial matching - prédiction par correspondance partielle) [8] a suscité un intérêt croissant des chercheurs dans le domaine du traitement du langage naturel, et a été utilisé pour résoudre divers problèmes dans ce domaine [12]. PPM présente une variante du mélange des probabilités, quand les probabilités obtenues dans les contextes de diverses longueurs sont unies dans une probabilité commune. Dans le cas général, la probabilité mélangée ***p(s)*** peut être calculée comme suit:

$$p(s) = \sum_{i=1}^{o} p'(c_i | c_{i-o}\, c_{i-o+1}\, c_{i-o+2} \ldots c_{i-1}) \qquad (1)$$

où $p'(c_i | c_{i-o}\, c_{i-o+1}\, c_{i-o+2} \ldots c_{i-1})$ – les probabilités du symbole courant, déterminé par le modèle pour tous les contextes, depuis le contexte maximum *o*.

Par exemple, la probabilité de la lettre « m » dans le contexte du mot « **algorithm** » dans le modèle PPM est calculée comme suit:

P_{PPM} ('***m***') = λ_5 P('***m***' | '***orith***') + λ_4 P('***m***' | '***rith***') +
+ λ_3 P('***m***' | '***ith***') + λ_2 P('***m***' | '***th***') + λ_1 P('***m***' | '***h***') +
+ λ_0 P('***m***') + λ_{-1} P('esc'), (2)

où λ_i (i = -1 ... 5) est le facteur de normalisation; 5 - contexte maximum; P('esc') - la probabilité d'abandon.

Dans [12], l'influence de la longueur du contexte sur la qualité du modèle a été analysée et on a observé que la longueur égale à 5 symboles précédents est optimale pour la compression des textes.

IV. L'APPLICATION DE PPM POUR RÉTABLIR LES SIGNES DIACRITIQUES

Dans [12] a été proposée l'application de la méthode PPM pour corriger les textes au niveau des lettres. Si nous supposons que les mots, qui devraient avoir des signes diacritiques, et ne les ont pas, sont mal orthographiés, alors nous pouvons utiliser l'algorithme de correction des textes pour corriger les erreurs de ce type. Pour la correction automatique des erreurs dans les textes est utilisé un tableau appelé « le tableau des remplaçants » dans lequel s'enregistrent des fragments de texte incorrects et les alternatives possibles de remplacement pour obtenir le texte correct.

La langue roumaine contient cinq signes diacritiques: ***ă, â, î, ş*** et ***ţ***. Dans le cas de restauration des signes diacritiques le tableau de remplacement est simple, il ne contient que quatre lettres (***a, i, t, s***) et les remplacements possibles sont (***ă, î, ţ, ş***). Pour vérifier la nécessité de remplacement l'entropie du fragment du texte pour la variante sans et avec les signes diacritiques est calculée. La variante avec une entropie minimale est supposée être correcte. Ce qui suit est l'algorithme de restauration des signes diacritiques dans le modèle PPM.

L'ALGORITHME de restauration des signes diacritiques à l'aide du modèle PPM:

Pour chaque lettre du texte $\mathbf{c_i}$:
  si $\mathbf{c_i}$ est la lettre ambigüe[1] (a, i, t, s):
    on calcule l'entropie $\boldsymbol{E_i}$ du segment de texte où est $\mathbf{c_i}$;
      remplacez $\mathbf{c_i}$ par la lettre avec signe diacritique $\mathbf{c'_i}$;
      calculez l'entropie $\boldsymbol{E'_i}$ du même segment de texte;
        si $\boldsymbol{E'_i < E_i}$ alors le remplacement est définitif;
  par contre, il reste la lettre initiale $\mathbf{c_i}$.

Les premières expériences pour le roumain sont décrites dans [1]. Elles contiennent la comparaison de 3 méthodes qui définissent le segment de texte comme suit: (1) seulement la lettre ambigüe; (2) le mot dans lequel se trouve la lettre ambigüe; (3) une fenêtre qui a la longueur 2n+1 (y compris des blancs) avec la lettre ambigüe au milieu. Chacune des trois méthodes décrites ont été utilisées avec des modèles PPM d'ordre différents: PPM d'ordre 3, 4, 5, etc. Les résultats obtenus démontrent que la longueur du contexte n'influe pas trop la qualité de restauration des signes diacritiques, ainsi dans les expériences suivantes nous avons utilisé le modèle PPM d'ordre 5. La troisième méthode est la meilleure et dans les expériences suivantes nous avons utilisé seulement cette méthode, donc on n'a pas besoin de diviser le texte en mots.

Nous avons effectué la prochaine série d'expériences sur la base du corpus JRC-Acquis [11] qui contient des textes en 22 langues considérées comme langues de travail officielles au Parlement européen. Pour les expériences nous avons choisi quatre langues, à savoir: le roumain, le français, l'italien et l'espagnol. Le corpus utilisé dans les expériences est constitué de 4400 documents pour chaque langue. Nous avons utilisé 6-parties de validation croisée (6-fold cross-validation), chaque partie contenant 730 documents). Ainsi, l'entrainement a été effectué chaque fois sur 3650 documents et testé sur 730. Dans les textes pour les tests les lettres avec des signes diacritiques sont remplacées par leurs paires sans aucun signe diacritique (par exemple, **é** par **e**, **à** par **a**, etc.)

La qualité de la restauration a été calculée au niveau des lettres et des mots. Le pourcentage présenté dans les expériences indique le nombre total de lettres/mots correctes après la restauration des diacritiques.

[1] Dans l'article, les auteurs ont tenu compte des rectifications orthographiques adoptées en 1990 par le Conseil Supérieur de la Langue Française et l'Académie Française (*Les rectifications de l'orthographe*. Document en ligne, consulté le 2011-28-01. http://www.academie-francaise.fr/langue/rectifications_1990.pdf).

La langue française utilise cinq signes diacritiques: les accents aigu, grave et circonflexe, le tréma et la cédille. Ils sont utilisés pour changer la forme, la valeur orthographique, la prononciation des mots. Par exemple, «un policier tue», ce n'est pas la même chose que «un policier tué».

Nous avons fait une étude statistique sur les mots dans le corpus qui contient 10745626 mots, dont 2146850 sont des mots accentués, qui constituent 20% de tous les mots du texte. Il est un peu plus faible que celui de la langue roumaine, mais le taux de mots accentués est élevé par rapport à d'autres langues.

Nous avons donc fait 6-parties de validation croisée (6-fold cross-validation) en utilisant le corpus français dans la méthode PPM d'ordre 5 basée sur le fragment, et nous avons obtenu les résultats suivants: le programme a restauré correctement 99,57% des lettres. Nous avons lancé le programme six fois et le résultat minimum a été 99,52% et 99,59% - le résultat maximum. Nous avons calculé le pourcentage au niveau des mots. Le programme a restauré correctement 98,01% des mots, le résultat minimum de 6 séries est de 97,52% et 98,22% pour le maximum. Les résultats sont présentés dans le tableau 1. Par rapport à la première expérience sur la langue roumaine, nous avons obtenu de meilleurs résultats. Cela peut s'expliquer par le plus petit nombre de mots accentués en français ce qui simplifie le processus de restauration diacritique. Une deuxième cause peut être le corpus qui est plus large, et peut contenir moins d'erreurs.

TABLE I. LES RESULTATS SUR LA LANGUE FRANÇAISE

La valeur moyenne sur les lettres	99,57%
Le résultat minimum pour les lettres	99,52%
Le résultat maximum pour les lettres	99,59%
La valeur moyenne sur les mots	98,01%
Le résultat minimum pour les mots	97,52%
Le résultat maximum pour les mots	98,22%

JRC-Acquis corpus contient aussi une partie roumaine et nous avons répété les expériences sur la langue roumaine dans le corpus donné. La méthodologie expérimentale a été identique à celle utilisée pour la langue française. Le programme a restauré correctement 99,43% des lettres et 97,32% des mots. Par rapport à la première expérience pour le roumain, nous avons obtenu de meilleurs résultats, ce qui signifie que le corpus a une influence majeure.

TABLE 2. LES RESULTATS SUR LA LANGUE ROUMAINE

La valeur moyenne sur les lettres	**99,43%**
Le résultat minimum pour les lettres	**99,39%**
Le résultat maximum pour les lettres	**99,47%**
La valeur moyenne sur les mots	**97,32%**
Le résultat minimum pour les mots	**97,08%**
Le résultat maximum pour les mots	**97,43%**

L'expérience suivante a été effectuée sur l'espagnol. Nous avons calculé les données statistiques d'apparition des signes diacritiques dans cette langue en utilisant le corpus du JRC-Acquis. Les lettres sans signes diacritiques font 97,79%, donc 2,21% des lettres ont des signes diacritiques. Au niveau des mots, 93% des mots dans le corpus n'ont pas de signes diacritiques. Par rapport au français et au roumain, les signes diacritiques sont utilisés moins fréquemment. Respectivement les résultats de restauration sont meilleurs. Nous avons répété exactement la même méthodologie que pour le français. Le programme a récupéré 99,86% des lettres et respectivement 99,52% de mots correctement. Nous avons lancé le programme six fois, et le résultat minimum a été 99,7%, le résultat maximum 99,92% au niveau des lettres. Au niveau des mots, le résultat minimum de 6 séries est de 98,26% et 99,6% pour le maximum.

TABLE 3. LES RESULTATS SUR LA LANGUE ESPAGNOL

La valeur moyenne sur les lettres	**99,86%**
Le résultat minimum pour les lettres	**99,7%**
Le résultat maximum pour les lettres	**99,92%**
La valeur moyenne sur les mots	**99,26%**
Le résultat minimum pour les mots	**98,52%**
Le résultat maximum pour les mots	**99,6%**

La dernière expérience a été effectuée pour la langue italienne. Au niveau des mots, 98,53% de mots ne contiennent pas de lettres avec des signes diacritiques, 1,47% des mots ont des signes diacritiques. Après l'exécution du programme 99,91% des lettres et respectivement 99,51% de mots sont corrects. Le résultat minimum a été 99,88% et le résultat maximum 99,94% au niveau des lettres. Au niveau des mots le résultat minimum de 6 séries est de 99,39% et 99,63% pour le maximum.

TABLE 4. LES RESULTATS SUR LA LANGUE ITALIENNE

La valeur moyenne sur les lettres	**99,91%**
Le résultat minimum pour les lettres	**99,88%**
Le résultat maximum pour les lettres	**99,94%**
La valeur moyenne sur les mots	**99,51%**
Le résultat minimum pour les mots	**99,39%**
Le résultat maximum pour les mots	**99,63%**

V. DISCUSSION

Nous avons analysé les erreurs obtenues par le programme dans le processus de restauration des signes diacritiques. Comme nous l'avons mentionné précédemment, la méthode appliquée est statistique et s'applique au niveau des caractères. Ainsi, elle vérifie chaque lettre ambigüe et décide d'insérer le diacritique pour cette lettre ou non. La statistique des lettres présentées dans les sections précédentes montre que le taux de lettres sans signes diacritiques est beaucoup plus élevé que celui de lettres avec diacritiques. Par exemple, le taux de la lettre « e » sans signes diacritiques dans les textes français est de 79,71% et avec tous les signes diacritiques possibles seulement 20,29%. Ainsi, le système vérifie toutes les lettres « e » dans le texte et fournit parfois des signes diacritiques supplémentaires là où ils sont inutiles. L'analyse des erreurs montre que le système ne restaure pas les signes diacritiques nécessaires plus souvent qu'il n'en insère par erreur. Une grande partie des erreurs sont commises au début des mots ou dans des mots courts. Le système se base sur les lettres précédentes et les lettres de début ont moins d'informations. Nous avons remarqué qu'une grande partie des erreurs sont commises dans les mots qui ne sont pas vraiment ambigus. Ces mots [13] sont

appelés U-mots (sans ambigüité, ne contenant qu'une seule forme de mots avec des signes diacritiques, et ne sont pas valables dans la langue si le signe diacritique manque). Ces mots peuvent être restaurés correctement et facilement en utilisant le dictionnaire. Parmi les mots qui ont été mal corrigés par notre programme presque 40% sont des U-mots, 8% sont des mots qui n'ont pas de signes diacritiques et 51% sont des mots ambigus. De nombreux mots ambigus contiennent les accentuations liés au contexte des mots (par exemple, parle - parlé). Ces types de cas seront traités dans l'avenir.

VI. CONCLUSION

Dans cet article nous présentons une méthode statistique de restauration de diacritiques dans des textes. Cette méthode est indépendante de la langue du texte. Elle nécessite un volume relativement petit de textes avec signes diacritiques et ne recourt pas à d'autres ressources lexicales et produits logiciels. En comparaison avec d'autres méthodes cette méthode a un grand nombre d'avantages:

- elle ne nécessite pas de ressources ou outils avancés spécifiques pour les langues traitées: des lexiques volumineux, analyseurs morphologiques ou syntactiques ;
- elle est appliquée au niveau des lettres, donc la méthode n'exige pas la segmentation du texte en mots (tokenization) ;
- la seule ressource requise est un corpus assez petit de textes corrects avec des signes diacritiques;
- elle est pratiquement indépendante de la langue, l'adaptation à une autre langue européenne est élémentaire ;
- la qualité de la restauration des diacritiques est au niveau des méthodes proposées par d'autres auteurs ;

L'algorithme est implémenté en langage Perl et peut être inclu dans un système de traitement du texte.

Cette méthode a été évaluée sur les langues roumaine, française, italienne et espagnole. À l'avenir, on prévoit de continuer les expériences notamment sur les langues européennes de l'Est dans lesquelles le nombre des diacritiques est élevé. De plus, on envisage de réaliser une version en ligne de ce système.

RÉFÉRENCES

[1] Bobicev, V. (2007) O altă metodă de restabilire a semnelor diacritice. Atelierul „Resurse lingvistice şi instrumente pentru prelucrarea limbii române", 179-188.

[2] Cleary, J.G., Witten, I.H., (1984). Data compression using adaptive coding and partial string matching. IEEE Transactions on Communications 32(4), 396-402.

[3] Daelemans, W., Zavrel, J., Van Der Sloot, K., Van Den Bosch, A., (2003). TiMBL: Tilburg memory based learner, version 5.0 reference guide. ILK Technical Report ILK 03-10, p.56.

[4] De Pauw, G., Wagacha, P. W., de Schryver, G-M. (2007). Automatic diacritic restoration for resource-scarce languages. Actes de Text, Speech and Dialogue, Tenth International Conference. 170-179.

[5] El-Beze, M., Merialdo, B., Rozeron, B., Derouault, A., (1994). Accentuation automatique des textes par des metodes probabilistes. Techniques et sciences informatique 16(6), 797-815.

[6] Galicia-Haro, S. N., Bolshakov, I. A., Gelbukh , A. F., (1999). A simple Spanish part of speech tagger for detection and correction of accentuation error. Actes de Second International Workshop on Text, Speech and Dialogue, 219-222.

[7] Mihalcea, R., Nastase, V., (2002). O metodă automată pentru inserarea diacriticelor în texte. "Limba Româna în Societatea Informationala - Societatea Cunoasterii", D. Tufis and F. G. Filip Editors, Academia Româna, 191-206.

[8] Moffat, A., (1990). Implementing the PPM data compression scheme. IEEE Transactions on Communications, 38, No. 11, 1917-1921.

[9] Nagy, G.N., Sabourin, M., (1998). Signes diacritiques: perdus et retrouvés. Actes de Colloque International Francophone sur l'Ecrit et le Document CIFED, 404-412.

[10] Simard, M. (1998). Automatic Insertion of Accents in French Texts. In Proceedings of EMNLP-3, Granada, Spain.

[11] Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., Varga, D., (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. Actes de 5th International Conference on Language Resources and Evaluation (LREC'2006), 2142-2147.

[12] Teahan, W. J., (1998). Modelling English Text. PhD thesis, University of Waikato, p. 243.

[13] Tufiş, D., Chiţu, A., (1999). Automatic Insertion of Diacritics in Romanian Texts. Actes de 5th International Workshop on Computational Lexicography COMPLEX, 185-194.

[14] Wagacha, P., De Pauw, G., Githinji, P., (2006). A grapheme-based approach for accent restoration in G˜ık˜uy˜u. Actes de Fifth International Conference on Language Resources and Evaluation, ELRA, 1937–1940

[15] Yarowsky, D., (1994). Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French. Actes de 32nd Annual Meeting of the Association for Computational Linguistics, 88-95.

[16] Yarowsky, D., (1999). Corpus-based techniques for restoring accents in Spanish and French texts. Natural Language Processing Using Very Large Corpora. Kluwer Academics Publisher, 99-120