Accentuation de mots inconnus : application au thesaurus biomédical MeSH*

Pierre Zweigenbaum, Natalia Grabar
DIAM — STIM/DSI, Assistance Publique – Hôpitaux de Paris
& Département de Biomathématiques, Université Paris 6
{pz,ngr}@biomath.jussieu.fr

Mots-clefs – Keywords

Réaccentuation, mots inconnus, étiquetage, langue de spécialité, médecine
Reaccenting, unknown words, tagging, specialized language, medicine

Résumé - Abstract

Certaines ressources textuelles ou terminologiques sont écrites sans signes diacritiques, ce qui freine leur utilisation pour le traitement automatique des langues. Dans un domaine spécialisé comme la médecine, il est fréquent que les mots rencontrés ne se trouvent pas dans les lexiques électroniques disponibles. Se pose alors la question de l'accentuation de mots inconnus : c'est le sujet de ce travail.

Nous proposons deux méthodes d'accentuation de mots inconnus fondées sur un apprentissage par observation des contextes d'occurrence des lettres à accentuer dans un ensemble de mots d'entraînement, l'une adaptée de l'étiquetage morphosyntaxique, l'autre adaptée d'une méthode d'apprentissage de règles morphologiques. Nous présentons des résultats expérimentaux pour la lettre *e* sur un thesaurus biomédical en français : le MeSH. Ces méthodes obtiennent une précision de 86 à 96 % (±4 %) pour un rappel allant de 72 à 86 %.

Some textual or terminological resources are written without diacritic marks, which hinders their use for natural language processing. Moreover, in a specialized domain such as medicine, all words are not always found in the available lexicons. The issue of accenting unknown words then arises. This is the subject of the present work.

We propose two accentuation methods which both rely on a learning process, based on the observation of the contexts of occurrence of the accentuable letters in a training corpus. One is adapted from a part-of-speech tagging method, the other from a method for learning morphological rules. We present experimental results for letter *e* on a French biomedical thesaurus: the MeSH. These methods obtain a precision which ranges from 86 to 96% (±4 %) and a recall from 72 to 86%.

*Ce travail prolonge une première expérience présentée au workshop « NLP in Biomedical Applications ».

1 Introduction

De nombreuses langues emploient l'alphabet dit latin avec des signes diacritiques : le français emploie ainsi quatre *e* accentués (*éèêë*) en plus de la forme non accentuée *e*. La prise en compte correcte de ces lettres accentuées au-delà de l'ASCII américain n'a pas été immédiate et générale. De nos jours, une panoplie d'encodages de caractères appropriés sont en usage quasi-universel, comme la famille ISO-latin ou Unicode. Cependant, certains textes et certaines ressources terminologiques sont encore, pour des raisons historiques, écrits en « typographie pauvre ». Par exemple, la version française du thesaurus MeSH (Medical Subject Headings, (INSERM, 2000), quelque 29 000 termes), est écrite en majuscules non accentuées[1]. C'est une source de difficulté lorsque l'on veut utiliser ces termes dans des interfaces en langue naturelle ou pour l'indexation automatique de documents textuels: un mot non accentué peut correspondre à plusieurs mots, produisant des ambiguïtés artificielles comme *cote* (*côte*, *côté*, *cote*).

Une façon simple de traiter ce problème est de supprimer tous les accents de toutes les interfaces... Cela simplifie l'appariement, mais accroît l'ambiguïté, dont les systèmes de traitement automatique des langues ont déjà leur content. L'un de nos objectifs, par ailleurs, est de construire des ressources langagières (lexiques, bases de connaissances morphologiques, etc.) pour le domaine médical (Zweigenbaum, 2001) et d'acquérir des connaissances linguistiques à partir de terminologies et de corpus, dont le MeSH (Grabar & Zweigenbaum, 1999). Meilleure est la qualité des données source, meilleures seront les chances d'apprendre des modèles de données linguistiques pertinents. Tous ces points militent en faveur de sources accentuées.

Nous avons donc entrepris de produire une version accentuée du MeSH français. Ce thesaurus comprend 19 971 termes et 9 151 synonymes, avec 21 475 formes de mots différentes. Sa réaccentuation manuelle complète n'est donc pas une petite tâche. Il serait bien sûr préférable de disposer d'une méthode qui génère automatiquement une forme accentuée de chaque mot, de l'appliquer au MeSH et de valider ses résultats. Une fois une telle méthode applicable, elle pourrait de plus s'appliquer à d'autres textes et terminologies. Et de fait, l'expérience de la préparation de corpus étiquetés ou arborés montre qu'il est généralement plus efficace pour quelqu'un de corriger une première annotation que d'en produire une à partir de zéro. Une accentuation de termes techniques effectuée directement à la main est par ailleurs quelquefois irrégulière, et ce type de méthode assistée peut aider à obtenir une accentuation plus stable.

L'équipe CISMeF du CHU de Rouen a déjà accentué plus de 5 500 termes MeSH, qui sont utilisés dans le catalogue CISMeF des sites médicaux francophones (Darmoni *et al.*, 2000) (`www.chu-rouen.fr/cismef`). D'une part, une quantité moindre de termes a besoin d'être accentuée. Cela produit d'autre part une base d'entraînement pour une procédure d'apprentissage. En fait, en commençant à accentuer manuellement des mots du MeSH, nous avons pu vérifier que des régularités d'accentuation de certaines parties de mots se retrouvaient dans un certain nombre de cas – par exemple, *reac* se réaccentue en *réac* dans *réaccentuation* ou *réaction*. Cela nous a encouragés à regarder plus loin dans cette direction, et à chercher des méthodes d'accentuation de termes non accentués. Cependant, les méthodes que nous avons trouvées dans la littérature ne traitent pas le cas des mots « inconnus », c'est-à-dire qui ne se trouvent pas dans le lexique utilisé par le système concerné. Malgré l'emploi de lexiques généraux et spécialisés, un nombre important des mots du MeSH sont dans ce cas[2]. On peut bien sûr avancer

[1] Il semble que ce soit une tradition tenace en Sciences de l'information.

[2] Citons par exemple *cachectine*, *cadherines*, *caenorhabditis*, *caeruleine*, *calcifiante*, *calcimedine*, *calcimedines*, *calcineurine*, *calciphylaxie*, *calcoaceticus*, *calelectrine*, *calelectrines*, *calendula*.

que la constitution d'un lexique plus complet devrait réduire la proportion de ces mots. Mais il s'agit pour la plupart de mots spécialisés et rares, dont nous avons eu du mal à trouver certains même dans un grand dictionnaire médical de référence (Garnier & Delamare, 1992). Il est donc pertinent de chercher à accentuer automatiquement ces mots inconnus.

Après une brève revue de méthodes existantes (section 2), nous abordons donc la question de la génération d'une forme accentuée pour des mots inconnus. Nous présentons deux méthodes (section 3) fondées sur un entraînement sur un ensemble de mots accentués. Nous montrons des résultats expérimentaux pour la lettre *e* sur le MeSH français (section 4) : de l'ordre de 86 % des mots contenant au moins un *e* sont correctement accentués (y compris correctement laissés sans accent) ; et la seconde méthode permet de délimiter une portion de l'ensemble de mots à accentuer (ici, 72–75 %) sur laquelle elle obtient une accentuation plus sûre (ici, de 95 %).

2 État de l'art

Des travaux antérieurs se sont attaqués à l'accentuation de textes, par exemple, (Spriet & El-Bèze, 1997), ou le projet Réacc (`www-rali.iro.umontreal.ca`, (Simard, 1998)). Ils traitent le cas où tous les mots possibles sont supposés connus : ils sont recensés dans le lexique. Le problème est alors de désambiguïser un mot non accentué lorsqu'il correspond à plusieurs formes du lexique – l'exemple *cote* de l'introduction. Cette désambiguïsation s'appuie sur les techniques de la désambiguïsation morphosyntaxique : si l'ambiguïté d'une occurrence de mot désaccentué est corrélée à une ambiguïté morphosyntaxique, l'examen du contexte permet la plupart du temps de déterminer la forme accentuée correcte de cette occurrence.

En revanche, nous n'avons trouvé aucune référence sur l'accentuation de mots *inconnus* : une méthode qui propose une version accentuée des mots hors du lexique. Nous avons tout d'abord cherché à éliminer de ce traitement les mots que l'on peut trouver dans des lexiques ou corpus accentués. Mais cette technique est limitée par la taille du corpus qui serait nécessaire pour que des mots « rares » y apparaissent, et par l'indisponibilité de lexiques électroniques spécialisés pour notre domaine. Nous avons alors cherché à mettre au point, par des techniques d'apprentissage, des règles d'accentuation qui puissent être appliquées à un mot non accentué pour générer une forme accentuée de ce mot. Nous les décrivons ci-dessous.

3 Accentuer des mots inconnus

3.1 Des mots « inconnus »

Le MeSH français a été brièvement présenté dans l'introduction ; nous travaillons avec sa version 2001 (il est mis à jour annuellement). La portion accentuée et mise en casse mixte par l'équipe CISMeF est celle de novembre 2001. Comme de nouvelles ressources sont régulièrement ajoutées dans CISMeF et peuvent utiliser des termes MeSH supplémentaires, un nombre plus grand de termes accentués est maintenant disponible. La liste des mots qui apparaissent dans les termes accentués de novembre 2001, soit 4 861 formes, nous sert de lexique de base. Nous avons retiré de cette liste les quelques « mots » contenant des chiffres, ceux ne contenant qu'une ou deux lettres (abréviations), et nous les avons mis en minuscules. Le lexique résultant compte 4 763 mots (4 756 différents une fois désaccentués). Ce lexique contient des mots sim-

ples : il ne cherche pas à recenser des unités polylexicales comme *infarctus du myocarde*, mais les découpe en *infarctus* et *myocarde*. En tout, on y trouve 6 670 occurrences de la lettre *e*, dont 1 789 accentuées. 1 522 mots contiennent au moins un *e* accentué ; 106 mots contiennent au moins une lettre accentuée autre que *e*.

Un mot est considéré comme inconnu lorsqu'il n'est pas listé dans notre lexique. Une première préoccupation consiste à ajouter davantage de mots connus dans ce lexique, dans l'espoir qu'ils finissent par couvrir la plupart des mots du MeSH. La question est alors de trouver des sources appropriées de mots supplémentaires. Nous avons utilisé pour cela plusieurs listes de mots spécialisés trouvées sur le Web (lexique sur le cancer, lexique médical général), ainsi que le lexique électronique de l'ABU (`abu.cnam.fr/DICO`), qui contient quelque 300,000 entrées pour le français « général ». Plusieurs corpus ont fourni des sources accentuées pour étendre ce lexique avec des mots médicaux (cardiologie, hématologie, réanimation, tirés de l'état courant du corpus CLEF (Habert *et al.*, 2001), et notices de médicaments). Le lexique résultant compte 276 445 formes différentes.

Après application de cette liste au MeSH, 7 407 mots restaient non reconnus. Nous avons également mis en minuscules les mots de cette liste, nous en avons retiré ceux ne contenant qu'une ou deux lettres, ceux contenant des chiffres et plus généralement n'importe quel caractère hors de l'ensemble des minuscules non accentuées (il restait quelques séparateurs et quelques mots qui étaient déjà accentués). La liste résultante compte 6 868 mots non accentués. Parmi ceux-ci, 5 188 comprennent au moins une fois la lettre *e* ; les exemples de la note 2 sont tirés de cette liste. Comme on peut s'en rendre compte, il ne s'agit pas en général de mots que l'on trouve dans un dictionnaire de langue générale. Ce sont ces mots qu'il s'agit d'accentuer.

3.2 Représenter le contexte d'une lettre à accentuer

Nous avons cherché à mettre au point des règles d'accentuation de lettres non accentuées fondées sur l'hypothèse suivante : le contexte dans lequel apparaît cette lettre (les lettres voisines) va permettre de décider de son accentuation.

Se pose alors la question de trouver une description appropriée du contexte d'une lettre *pivot* dans un mot, par exemple la lettre *é* dans *excisée*. Une solution pourrait être de fixer un nombre de lettres autour du pivot. Des séquences de deux ou trois lettres (bigrammes, trigrammes) sont souvent choisies comme base pour des méthodes d'apprentissage (par exemple, l'étiquetage probabiliste (Weischedel *et al.*, 1993)). Mais il semble qu'un contexte plus large soit souvent nécessaire ; par exemple, *eme* s'accentue généralement en *ème* à la fin d'un mot (*emphysème*, *quatrième*, *problème*), alors qu'il s'« accentue » en *eme* lorsqu'il est suivi d'un *n* (*arrachement*, *infarcissement*, *sexuellement*). De plus, le contexte pertinent autour d'une lettre pivot peut s'étendre à la fois à gauche et à droite du pivot ; par exemple, pour la série des *e*, être suivi d'un *x* est un indice fort d'un *e* sans accent (*exemple*, *flexion*), comme l'est le fait de suivre un *é* (*laryngée*, *calcanéen*, *caséeux* ; seuls quelques mots français comme *créé* ont deux *é* à la suite).

Nous avons repris deux façons de représenter le contexte d'une lettre. D'une part, la représentation employée par l'étiqueteur de Brill (Brill, 1995) décrit des contraintes sur le voisinage d'un mot dans une phrase ; nous l'avons transposée au voisinage d'une lettre dans un mot (section 3.3). D'autre part, une notion de *contexte mixte* a été présentée dans (Theron & Cloete, 1997) pour apprendre des règles morphologiques à partir d'exemples. Cette notion permet de représenter des contextes à longueur variable qui s'étendent des deux côtés d'une lettre pivot

(section 3.4). Elle prend la forme de chaînes de caractères repliées, que l'on peut également mettre sous forme de transducteurs à états finis.

3.3 L'accentuation comme un étiquetage contextuel

Cette première méthode est fondée sur l'emploi d'un étiqueteur morphosyntaxique : celui de Brill (Brill, 1995). Nous considérons chaque mot comme une « phrase de lettres » : chaque lettre constitue un mot, et la suite des lettres d'un mot constitue une phrase. L'étiquette d'une lettre est la forme accentuée de cette lettre, ou la même lettre si celle-ci n'est pas accentuée. Par exemple, pour le mot *endometre* à accentuer en *endomètre*, la « phrase » est *e/e n/n d/d o/o m/m e/è t/t r/r e/e* (dans le format de l'étiqueteur de Brill). L'apprentissage habituel de l'étiqueteur permet alors d'apprendre des règles contextuelles d'accentuation, dont les premières sont indiquées sur le tableau 1. Par exemple, lors de l'accentuation du mot *flexion*, la règle (1) s'applique d'abord

Format Brill	Glose	Format Brill	Glose
(1) e é NEXT2TAG i	*e.i* ⇒ e → é	(2) e é NEXT1OR2TAG o	*e.?o* ⇒ e → é
(3) e é NEXT1OR2TAG a	*e.?a* ⇒ e → é	(4) e é NEXT1OR2WD e	*e.?e* ⇒ e → é
(5) e é NEXT2TAG h	*e.h* ⇒ e → é	(6) é è NEXTBIGRAM n e	*éne* ⇒ é → è
(7) é e NEXTBIGRAM m e	*éme* ⇒ é → e	(8) e é NEXTBIGRAM t r	*etr* ⇒ e → é
(9) é e NEXT1OR2OR3TAG x	*é.?.?x* ⇒ é → e	(10) e é NEXT1OR2TAG y	*e.?y* ⇒ e → é
(11) e é NEXT2TAG u	*e.u* ⇒ e → é	(12) e é SURROUNDTAG t i	*tei* ⇒ e → é
(13) é è NEXTBIGRAM s e	*ése* ⇒ é → è	(14) é e NEXT1OR2OR3TAG x (bis)	*é.?.?x* ⇒ é → e
(15) é e PREVTAG o	*oé* ⇒ é → e	(16) é e PREVBIGRAM y p	*ypé* ⇒ é → e

Table 1: Règles de correction d'accentuation.

et accentue le *e* pour donner *fléxion* (comme dans ...*émie*). La règle (9) s'applique ensuite pour corriger cette accentuation devant un *x*, ce qui redonne *flexion*. Ces règles correspondent à une représentation du contexte d'occurrence d'une lettre. Cette représentation est mixte (contextes gauche et droite peuvent être combinés, par exemple dans SURROUNDTAG), et peut aller jusqu'à une distance de trois lettres à gauche et à droite, mais avec des combinaisons contraintes. Contrairement à la méthode suivante, il s'agit ici de règles successives de correction de l'accentuation initiale la plus probable, qui consiste à n'accentuer aucun *e*.

3.4 Le voisinage d'une lettre en une chaîne de caractères

La représentation par un *contexte mixte* employée dans (Theron & Cloete, 1997) a l'intérêt de ne pas mettre de limite au nombre de caractères qui peuvent être pris en compte. Elle replie les lettres d'un mot autour de la lettre pivot considérée, en collectant alternativement la lettre suivante sur la droite puis sur la gauche, jusqu'à ce que les deux extrémités du mot soient atteintes. Ces extrémités sont marquées par des symboles spéciaux (ici, ^ pour le début du mot, et $ pour la fin du mot)[3]. Par exemple, le premier *e* de *excisée* est représenté par le contexte mixte de la colonne de droite de la première ligne du tableau 2. La colonne de gauche montre l'ordre dans lequel les lettres du mot sont énumérées. Les deux lignes suivantes indiquent les représentations en contexte mixte pour les deux autres *e* de ce mot.

[3]Theron & Cloete répètent de plus un symbole « hors-mot » lorsqu'une extrémité a été atteinte mais pas l'autre.

Mot									Contexte mixte=Cible
^	e	x	c	i	s	é	e	$	x ^ c i s e e $= e
2	.	1	3	4	5	6	7	8	
^	e	x	c	i	s	é	e	$	e s $ i c x e ^= é
8	7	6	5	4	2	.	1	3	
^	e	x	c	i	s	é	e	$	$ e s i c x e ^= e
8	7	6	5	4	3	2	.	1	

Table 2: Un mot et les représentations en contexte mixte de ses *e*.

Chacun de ces contextes est désaccentué, car il a pour vocation d'être comparé à des représentations de mots non accentués. La forme originale de la lettre pivot est associée au contexte sous forme d'une « sortie » (nous utilisons le symbole « = » pour noter cette sortie). Chaque contexte est donc converti en un transducteur : l'entrée est le contexte mixte autour d'une lettre pivot, et la sortie est la lettre (accentuée ou pas) appropriée parmi les lettres possibles.

L'étape suivante consiste à déterminer des contextes discriminants minimaux. Pour cela, nous fusionnons tous les transducteurs produits sur l'ensemble de mots d'entraînement (opération OU) en mettant en facteur leurs préfixes communs en un arbre lexicographique. Nous obtenons ainsi un transducteur déterministe qui représente exactement l'ensemble d'apprentissage. Nous calculons ensuite, pour chaque état de ce transducteur et pour chaque sortie possible (lettre de l'ensemble *eéèêë*) atteignable à partir de cet état, le nombre de chemins qui mènent de cet état à cette sortie. Nous appelons *non ambigu* un état dont tous les chemins sortants mènent à la même sortie. Dans ce cas, pour nos besoins, ces chemins peuvent être remplacés par un raccourci vers cette sortie commune. Cela revient à généraliser l'ensemble des contextes en les remplaçant par un ensemble de contextes minimaux discriminants.

Étant donné un mot à accentuer, la première étape est de représenter le contexte de chacune de ses lettres pivot. Par exemple, le mot *réel* contient deux *e* dont les contextes mixtes sont *erl^$* et *le$r^*. Le transducteur est appliqué à chaque contexte pour trouver le plus long chemin partant de son état de départ qui corresponde à un préfixe de ce contexte. Si ce chemin mène à un état de sortie, la sortie est la forme accentuée proposée pour la lettre pivot. Si l'appariement se termine plus tôt, nous sommes en présence d'une ambiguïté : plusieurs états de sortie peuvent être atteints, et nous ne prenons pas de décision[4]. Le nombre d'exemples de l'ensemble d'apprentissage qui ont pris le même chemin peut donner une évaluation du *support* (niveau de confiance) étayant un état de sortie donné (une décision d'accentuation). La procédure d'accentuation peut choisir de prendre une décision d'accentuation seulement lorsque le support pour cette décision est au-dessus d'un certain seuil.

Le tableau 3 montre quelques contextes discriminants minimaux appris sur les mots accentués du MeSH. Les trois premiers permettent d'accentuer les trois *e* de *excisée*[5]. Notons que d'autres représentations des contextes pourraient être employées. Nous avons examiné des *contextes droits* (une chaîne de longueur variable de lettres contiguës situées à droite du pivot) et des *contextes gauches* (*idem* à gauche) ; chacun donne des résultats inférieurs aux contextes mixtes.

[4]Notons qu'il serait possible dans ce dernier cas de prendre en compte les fréquences relatives d'occurrence de ces chemins alternatifs dans l'ensemble d'apprentissage, et de prendre des décisions probabilistes sur cette base (Zweigenbaum & Grabar, 2002). Dans les expériences présentées ici, seuls les cas non ambigus sont traités.

[5]On remarquera que le contexte -*ée*- n'est pas proposé ; en effet, l'ensemble d'entraînement contient des mots anglais comme *feed* qui le rendent ambigu.

Contexte	Support	Déplié	Exemples
x=e	68	*-ex-*	^exanthème$, ^anorexie$, ^cortex$
es=é	11	*-sée-*	^nausée$, ^médicalisée$
$e=e	53	*-ée*	^aménorrhée$, ^apnée$, ^assistée$
mho^=é	22	^*hémo-*	^hémodiafiltration$

Table 3: Des contextes discriminants minimaux.

3.5 Évaluation

Nous avons entraîné les deux méthodes sur deux ensembles d'entraînement : les 4 763 mots de la partie accentuée du MeSH (section 3.1) et les 54 291 lemmes du lexique ABU. Pour vérifier la validité des règles, nous les avons appliquées aux mêmes mots du MeSH désaccentués (4 756). Cela permet de mesurer un taux de rappel (mots correctement accentués / mots à accentuer) et de précision (mots correctement accentués / mots accentués). La méthode des contextes a été appliquée avec une série de seuils différents (0, 2, 5, 10, 20, 50). On s'attend à ce qu'un seuil plus haut améliore la précision au détriment du rappel.

On a indiqué que la méthode des contextes prend individuellement pour chaque lettre une décision d'accentuation. Elle sait donc pour quelle proportion des *e* d'un mot elle a pu décider. Nous répartissons alors les mots traités en trois classes : décision pour toutes les lettres *[complète]*, décision pour une partie des lettres *[partielle]*, décision pour *[aucune]* des lettres. Nous verrons que la précision est nettement meilleure sur la classe *[complète]*. Une stratégie possible est alors de n'accepter que les résultats pour cette classe, ce qui augmente la précision mais diminue le rappel. Nous donnons donc aussi une mesure individuelle pour cette classe.

Nous avons enfin appliqué les deux méthodes sur les 5 188 mots « inconnus » accentuables du MeSH (section 3.1). Aucun étalon n'est disponible pour valider cette accentuation : une validation humaine était donc nécessaire. Pour l'effectuer, nous avons prélevé au hasard des échantillons dans chacune des trois classes *[complète]*, *[partielle]* et *[aucune]*, et reconstitué un échantillon global formé de mots de ces trois classes au prorata de leurs tailles relatives. Ces échantillons ont été revus par l'équipe CISMeF. Du fait de l'échantillonnage, les mesures de précision doivent ici inclure un intervalle de confiance. Nous nous sommes servis de cette estimation de la précision pour calculer une estimation du rappel attendu pour chaque classe (taille relative de la classe $\times$ précision pour cette classe).

Les programmes ont été écrits en `Perl5`. Ils comprennent un paquetage de manipulation d'arbres lexicographiques que nous avons écrit en étendant le paquetage `Tree::Trie` en ligne sur le « Comprehensive Perl Archive Network » (`www.cpan.org`).

4 Résultats

L'étiqueteur de Brill apprend 80 (MeSH) ou 208 (ABU) règles contextuelles. La méthode des contextes apprend 1 740 (MeSH) ou 15 925 (ABU) contextes minimaux au seuil 0.

Le tableau 4 résume les résultats de la vérification sur les mots d'entraînement « MeSH » : la précision globale est très bonne pour Brill (89–91 %). Pour la méthode des contextes entraînée sur MeSH, elle est excellente au seuil 0 avec les deux ensembles d'apprentissage (93–99,6 %),

mais elle décroît lorsque le seuil augmente. Les chiffres de la classe *[complète]* peuvent expliquer ce fait : l'accentuation est parfaite pour cette classe, mais sa taille décroît lorsque le seuil croît ; les classes *[partielle]* et *[aucune]* voient donc leur taille augmenter, et ont chacune une précision moindre. Entraînée sur ABU, la méthode des contextes donne également de très bons résultats globaux au seuil 0, qui dès le seuil 2 sont meilleurs que ceux pour MeSH ; par ailleurs, la précision pour la classe *[complète]* est excellente (96 %) et augmente, elle, avec le seuil ; mais ici aussi, son rappel diminue lorsque ce seuil augmente.

Apprentissage		Brill	Méthode des contextes					
seuil :		–	0	2	5	10	20	50
MeSH	P	0,91	0,996	0,87	0,80	0,74	0,70	0,68
(globale)	R	0,92	0,995	0,87	0,80	0,74	0,70	0,68
MeSH	P		1	1	1	1	1	1
[complète]	R		0,84	0,57	0,41	0,28	0,18	0,10
ABU	P	0,89	0,93	0,89	0,84	0,79	0,75	0,71
(globale)	R	0,89	0,93	0,89	0,84	0,79	0,75	0,71
ABU	P		0,96	0,985	0,990	0,990	0,994	0,998
[complète]	R		0,70	0,52	0,39	0,29	0,19	0,10

Table 4: Test sur la partie accentuée du MeSH.

Le seuillage ne nous semble donc pas pertinent ; nous ne l'appliquerons pas par la suite. On observe que l'entraînement sur le MeSH, s'il obtient généralement des résultats un peu meilleurs qu'ABU, ne s'en distingue pas fortement. Il aurait donc probablement été possible d'employer cette méthode d'accentuation, entraînée sur ABU, pour préparer l'accentuation de cette partie du MeSH. Nous avons effectué les mêmes tests avec des contextes droits (précision moyenne 76 %) et gauches (70 %). Leur accentuation *[complète]* est elle aussi précise (93 et 94 % au seuil 0), mais avec un rappel très faible (3 et 22 %). Ces deux méthodes ont donc été délaissées.

L'accentuation des 5 188 mots « inconnus » du MeSH donne les résultats indiqués sur le tableau 5 (seuil 0). La colonne de droite (Base) indique la proportion de mots ne contenant aucun accent. Cette proportion correspond à la précision qu'aurait une procédure choisissant systématiquement de laisser tous les *e* non accentués.

Décision	Entraîn.t	Nb	% total	Éch.	Cor.	Précision±IC	Rappel±IC	Base
Brill	MeSH	5 188	100	250	215	0,86±0,04	0,86±0,04	0,68
Brill	ABU	5 188	100	250	215	0,86±0,04	0,86±0,04	
total	MeSH	5 188	100	250	215	0,86±0,04	0,86±0,04	0,71
total	ABU	5 188	100	250	206	0,82±0,05	0,82±0,05	
[complète]	MeSH	3 898	75	180	172	0,96±0,03	0,72±0,02	0,74
[complète]	ABU	3 734	72	142	135	0,95±0,04	0,68±0,03	
[partielle]	MeSH	784	15	86	39	0,45±0,11	0,07±0,02	0,40–5
[partielle]	ABU	823	16	38	21	0,55±0,16	0,09±0,03	
[aucune]	MeSH	506	10	71	64	0,90±0,07	0,09±0,01	0,90–6
[aucune]	ABU	631	12	20	19	0,95±0,10	0,12±0,01	

Table 5: Accentuation des mots inconnus (seuil = 0, précision sur échantillons Éch.).

On constate que les résultats globaux de l'étiqueteur de Brill (MeSH ou ABU) et de la méthode des contextes (MeSH) sur l'échantillon de test sont identiques – les résultats sont un peu moins bons pour la méthode des contextes entraînée sur ABU. Si l'on distingue les trois classes de décisions, les résultats sont en revanche très tranchés. La précision est excellente (95–96 %) lorsque la décision est complète, mauvaise (45–55 %) lorsque la décision est partielle, et bonne de nouveau (90–95 %) lorsqu'aucune décision n'a pu être prise. On peut donc obtenir une précision de 95 % pour un rappel de 72 % (*[complète]*), ou encore une précision à peine moins bonne ($\sim$93 %) pour un rappel de 81 % (*[complète]* + *[aucune]*).

Sur la classe *[partielle]*, Brill obtient une précision de 65$\pm$10 % (MeSH, hors tableau), nettement meilleure que celle de la méthode des contextes sur cette classe. Une combinaison avantageuse consisterait à utiliser d'abord la méthode des contextes, découpant l'ensemble de mots en trois classes ; à conserver ses résultats pour les classes *[complète]* et *[aucune]* ; et à appliquer l'accentuation par Brill sur la classe *[partielle]*. On pourrait en attendre un gain en précision globale de l'ordre de 3 % (MeSH : +20 % $\times$ proportion de *[partielle]* = 0,15).

5 Bilan

Les méthodes d'accentuation présentées ici permettent d'obtenir des performances absolues honorables. Il faut également juger de leur utilité dans le cadre d'une tâche d'assistance à l'accentuation humaine. La stratégie de combinaison des deux méthodes proposée dans la section précédente devrait faciliter le travail de l'accentueur-correcteur humain : la correction d'une classe de mots (*[complète]*, ici, près des trois quarts) ne contenant potentiellement que peu d'erreurs ($\sim$5 %), d'une classe de mots (*[aucune]*, $\sim$12 %) laissés non accentués et dont probablement peu sont accentuables ($\sim$5 %), et la correction d'une classe de mots (*[partielle]*, $\sim$15 %) dont on s'attend à ce qu'elle contienne davantage d'erreurs d'accentuation ($\sim$40 %). Notons que l'évaluation a été faite sur mots entiers. Une autre évaluation serait possible en distinguant l'accentuation individuelle de chaque *e*; ces décisions d'accentuation individuelles pourraient être rendues visibles dans une interface de validation.

Nous n'avons pas cherché à prendre en compte un découpage des mots en morphèmes. Ce découpage serait sans doute pertinent pour une partie des mots (par exemple, endo*mètre*, thermo*mètre*), et pourrait peut-être prédire avec plus de certitude l'accentuation de ces mots, sur la base des morphèmes ainsi segmentés. Certains des contextes appris sont certes liés à ces morphèmes (par exemple, *^hémo-*, vu plus haut, ou *tmh^yl$=é* pour le mot *méthyl*). Néanmoins, un contexte incluant un marqueur de début de mot ne permet pas lui-même d'accentuer un morphème en milieu de mot (par exemple, *phenylmethylsulfonyle*). Un marquage des frontières de morphèmes et un apprentissage de contextes s'arrêtant à ces frontières fonctionnerait dans ce cas, et s'appliquerait sans doute de façon plus générale. D'ailleurs, il semble qu'une part importante de la classe *[partielle]* soient des mots composés longs comme celui de cet exemple.

Une partie des mots mal accentués sont des mots anglais (*académic*, *cléavage*) qui se trouvent dans certains termes du MeSH français. On pourrait tenter de les repérer par l'emploi d'un lexique anglais. Mais ce n'est pas si simple dans le contexte où nous opérons, car un certain nombre de mots anglais ont une graphie identique à celle de mots français désaccentués. Enfin, nous avons traité seulement la lettre *e*. Il serait sans doute hâtif de généraliser les résultats obtenus ici aux autres lettres accentuées (*aiouç*). En effet, la régularité observée pour le *e* est sans doute liée à la traduction graphémique de règles phonologiques. La distribution des

lettres *âîïç* et peut-être encore davantage des *ùäüö* risque d'être plus idiosyncrasique et liée à des lexèmes particuliers.

Ceci étant posé, nous pensons que ces méthodes devraient aider à réduire substantiellement le temps humain nécessaire à accentuer non seulement le MeSH, mais aussi d'autres ressources. Par exemple, la base de connaissances ADM en ligne à l'Université de Rennes (Seka *et al.*, 1997) est un autre corpus médical précieux qui reste écrit en majuscules non accentuées.

Remerciements

Nous remercions Magaly Douyère, Benoît Thirion et Stéfan Darmoni, de l'équipe CISMeF, d'avoir mis à notre disposition les termes accentués du MeSH, et pour leur aide dans la validation des résultats de l'accentuation.

Références

BRILL E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, **21**(4), 543–565.

DARMONI S. J., LEROY J.-P., THIRION B., BAUDIC F., DOUYÈRE M. & PIOT J. (2000). CISMeF: a structured health resource guide. *Methods Inf Med*, **39**(1), 30–35.

GARNIER M. & DELAMARE V. (1992). *Dictionnaire des Termes de Médecine*. Paris: Maloine.

GRABAR N. & ZWEIGENBAUM P. (1999). Acquisition automatique de connaissances morphologiques sur le vocabulaire médical. In P. AMSILI, Ed., *Actes de TALN 1999 (Traitement automatique des langues naturelles)*, p. 175–184, Cargèse: ATALA.

HABERT B., GRABAR N., JACQUEMART P. & ZWEIGENBAUM P. (2001). Building a text corpus for representing the variety of medical language. In *Corpus Linguistics 2001*, Lancaster.

INSERM (2000). *Thésaurus Biomédical Français/Anglais*. Institut National de la Santé et de la Recherche Médicale, Paris.

SEKA L., COURTIN C. & LE BEUX P. (1997). ADM-INDEX: an automated system for indexing and retrieval of medical texts. In *Stud Health Technol Inform*, volume 43 Pt A, p. 406–410: Reidel.

SIMARD M. (1998). Automatic insertion of accents in French text. In *Proceedings of the Third Conference on Empirical Methods in Natural Language Processing*, Grenade.

SPRIET T. & EL-BÈZE M. (1997). Réaccentuation automatique de textes. In *FRACTAL 97*, Besançon.

THERON P. & CLOETE I. (1997). Automatic acquisition of two-level morphological rules. In R. GRISHMAN, Ed., *Proceedings of the Fifth Conference on Applied Natural Language Processing*, p. 103–110, Washington, DC: ACL.

WEISCHEDEL R., MEETER M., SCHWARTZ R., RAMSHAW L. & PALMUCCI J. (1993). Coping with ambiguity and unknown words through probabilistic models. *Computational Linguistics*, **19**(2), 359–382. Special Issue on Using Large Corpora: II.

ZWEIGENBAUM P. (2001). Ressources pour le domaine médical : terminologies, lexiques et corpus médicaux. *Lettre de l'ELRA*, **6**(4), 8–11.

ZWEIGENBAUM P. & GRABAR N. (2002). Accenting unknown words in a specialized language. In *ACL Workshop Natural Language Processing in the Biomedical Domain*, Philadelphie: ACL. *À paraître*.