

## Quelques problèmes spécifiques au français en AIML : la graphie

Soumis par Philippe YONNET

18-05-2008

Dernière mise à jour : 18-05-2008

L'écriture du français pose quelques problèmes spécifiques lorsque l'on crée des catégories en AIML. L'essentiel de ces problèmes provient de l'existence des caractères accentués, de l'élosion (les problèmes liés à la syntaxe seront abordés dans d'autres articles). Les caractères accentués

Le français utilise cinq signes dits "diacritiques" :

- l'accent aigu
- l'accent grave
- l'accent circonflexe
- le tréma
- la cédille

Le problème, c'est que l'usage veut que ces signes ne soient pas utilisés dans les majuscules (par erreur parait il car les majuscules DOIVENT être accentuées). Par ailleurs, de nombreux français, même bons en orthographe, commettent des erreurs d'accents. Enfin, nombreux sont ceux qui omettent de taper les accents dans les champs de recherche sur des applications internet, car il est fréquent que les accents soient tout bonnement ignorés.

Le problème, c'est que ce n'est pas le cas dans toutes les implémentations du langage. Par exemple, pour le programme Z (alias Pandorabots), "génération" ne matche pas avec "generation".

Au contraire, le programme E (implémentation en php) le permet par un paramétrage de startup.xml

Par prudence, il faudra donc générer les variantes des patterns avec des accents et sans les accents, pour permettre aux catégories de fonctionner sur toutes les implémentations.

Quelques bizarreries du français :

### Le tréma

le tréma ne se rencontre que dans les mots suivants :

- boësse, boëte, canoë, foëne, maërl, moëre, Azraël, Gaël, Ismaël, Israël, Joël, Judicaël, Michaël, Nathanaël, Noël, Raphaël, Staël, Laëtitia. Les mots goëland et poëme de graphie archaïque peuvent exister dans le langage poétique ;
- aïeul, ambiguïté\*, amuissement, stoïle, naïf, païen, pagaïe, baïonnette, coïncider, stoïque, archaïque, haïr, ouïe, ouïr, astéroïde, maïs, paranoïa, voltaïque, laïc, Loïc ;
- aiguë\*, ambiguë\*, béguë\*, capharnaüm, ciguë\*, Ésaü, Emmaüs, bisaiguë\*, crapaüter (variante de crapahuter), contiguë\*, exiguë\*.

Les rectifications de 1990 font remonter le tréma sur la voyelle effectivement prononcée. Les mots modifiés sont signalés ici par une astérisque. Pour ces mots il y'a donc plusieurs graphies possibles.

L'accent circonflexe apparait sur a, e, i, o et u.

L'accent grave n'apparait que sur a, et e (et sur u uniquement dans "où")

L'accent aigu n'apparait que sur e

La cédille sur c

=> il faut prévoir ces variantes dans notre générateur de patterns automatique

Nota bene : lorsque la variante avec accent correspond à une écriture diacritique (c'est à dire que l'accent sert à distinguer deux mots au sens ou à l'utilisation différente, comme ou et où, sur et sûr) il sera possible de créer une catégorie beaucoup plus intelligente. Il est donc intelligent de créer un dictionnaire spécial de ces formes diacritiques pour enrichir la génération automatique de catégories. L'élosion

En français, le e muet en fin de mot est remplacé par une apostrophe s'il est suivi d'une voyelle. Les articles comme le,

les pronoms comme je, de, me, te, se, les adverbes comme ne, le mot que, sont donc régulièrement élidés et s'écrivent l', m', t', s'.

L'élision s'étend à d'autres cas quand la prononciation de deux voyelles consécutives devient gênante : c'est vrai avec "la" (exemple la élision devient l'élision), avec du, avec si ("si il" devient s'il). Ces cas sont beaucoup plus rares.

Bizarreries et cas particuliers

Certains mots débutant par une voyelle ne peuvent être précédés d'un autre mot élidé. C'est vrai des mots commençant par un h aspiré. Par exemple, haricot.

Mais il y'a d'autres exceptions : onze, et huit

D'autres mots entraînent des élisions avec des règles particulières :

Entre, Jusque, Lorsque, Parce que, Presque, Puisque, Quelque, Quoique Par ailleurs des formes de langage parlé ou des formes "familières" se transcrivent par des élisions.

Problème, toutes les implémentations ne gèrent pas forcément les élisions correctement. Le programme E (php) a tendance à remplacer les apostrophes par rien, ce qui crée des catastrophes. Il faut donc modifier le programme, ou plus simplement, substituer toutes les formes d'élision possible par la séquence sans apostrophe. (remplacer "l'é" par "l é"). Cela se fait en modifiant le fichier startup.xml.

Le programme Z (Pandorabots) comme d'autres implémentations gèrent normalement les élisions (l'apostrophe est remplacée par un espace). Donc il n'est pas utile de créer les variantes avec et sans élision. Par contre, la disparition de l'apostrophe rend l'identification de la forme élidée difficile (on se retrouve avec des l s t j isolés dans les patterns).

Pour générer des catégories automatiquement de manière intelligente, il sera donc intéressant de reconnaître les formes élidées dans les patterns pour en tirer des informations utiles comme la personne (j' m' = 1ère personne, t' deuxième etc...).

Nous verrons dans un deuxième article les problèmes de syntaxe, comme l'existence de quasi mots vides (pronoms, articles), la multiplication des morphèmes, et les problèmes posés par le tour interrogatif, les subordonnées, et les différences entre langage parlé, et langage écrit.