Real-time Automatic Insertion of Accents in French Text

Michel Simard *Alexandre Deslauriers*

Laboratoire de Recherche Appliquée en Linguistique Informatique
Université de Montréal, Québec, Canada

{simardm, deslaura}@iro.umontreal.ca

1. Contexte

Le traitement automatique du langage naturel est une discipline à la frontière de la linguistique, de l'informatique et de l'intelligence artificielle. Elle concerne l'application de programmes et techniques informatiques à tous les aspects du langage humain.

L'analyse Morpho-Syntaxique étudie d'une part la forme (morphologie), et d'autre part la fonction (syntaxe) d'un élément du discours (d'où le nom "analyse en partie du discours" (ou POS tagging)). Le but de cette analyse est d'attribuer une étiquette morpho-syntaxique à chaque élément du texte. L'*analyse morpho-syntaxique* est, depuis plusieurs années, un problème résolu raisonnablement ($>$ 95%) par des méthodes probabilistes [4].

La réaccentuation de texte est un sous problème du domaine de ré-introduction de diacritiques dans les textes qui n'en sont pas pourvus. Les diacritiques sont des signes accompagnant les lettres et permettant de modifier leur valeur phonétique ou de lever des ambigüités entre homographes.

2. Motivation

Provenance des textes non accentués

L'absence de caractères accentués dans les textes Français est principalement due à :

- L'utilisation d'encodages (représentation informatique d'un jeu de caractères d'une langue naturelle) mals apropriés. ex : ascii 7-bit pour lire de l'iso-latin1.
- L'utilisation d'applications simplificatrices supprimant les caractères accentués.
- L'utilisation de claviers étrangers par des français rendant l'utilisation de diacritiques très laborieuse.

Cependant, certaines applications de traitement du langage naturel nécessitent des différenciations entre caractères non accentués et accentués comme, par exemple, la traduction automatique ou la synthèse vocale (ex : prononciation de "même" vs. "mémé").

Les (relativement) récents efforts de normalisation tels *Mime* et surtout *Unicode* sont apparus pour palier ces problèmes mais leur utilisation, bien qu'en forte augmentation, n'est pas systématique.

Etude du problème

En français, 85% des mots n'ont pas d'accents et 10% des mots sont non ambigus (un seul mot accentué correspondant au mot non-accentué) et peuvent donc être ré-accentués immédiatement à l'aide d'un dictionnaire.

Les 5% restant représentent des mots ambigus (plusieurs mots peuvent être associés au mot non-accentué). On peut, pour ces mots, calculer une probabilité à priori (approche fréquentiste sur un corpus accentué) et limiter ainsi les erreurs à 3% des mots du français.

Une méthode standard pour la résolution de ces derniers 3% serait la réalisation d'un modèle de langage *n-grammes* (qui répertorie les fréquences d'apparition des suites de *n-mots*). Cependant, 99.62% des mots ambigus peuvent être retrouvés par le biais de leur étiquette morpho-syntaxique [1]. On préfèrera alors réaliser un étiquetteur morpho-syntaxique statistique qui nécessite un corpus d'apprentissage moins important et utilise des modèles moins volumineux que la méthode des n-grammes.

3. Contribution

Les modèles de Markov cachés (ou HMM) permettent d'apprendre des appariements entre des séquences de mots et des séquences d'étiquettes de manière statistique.

Dans le cadre de cet article nous avons utilisé un modèle d'ordre 2 défini par :

- le vecteur π donnant la probabilité de commencer par une certaine étiquette
- la matrice de transition $\mathcal{A}$ donnant la probabilité de passer d'une étiquette t_{i-1} à une étiquette t_i : $P(t_i|t_{i-1})$
- la matrice d'observation $\mathcal{B}$ donnant la probabilité de voir un mot w_i avec une étiquette t_i : $P(w_i|t_i)$.

L'apprentissage d'un modèle revient alors à un calcul de fréquence pour ces trois paramètres sur un large corpus annoté et à une ré estimation de ceux-ci par des appels successifs à l'algorithme de Baum-Welch[5].

De plus, afin de limiter la combinatoire, nous avons :

- Limité les choix possibles d'étiquettes au moyen d'un lexique du Français.
- Divisé le texte en segments suivant un seuil $\mathcal{S}$.

Le seuillage est réalisé en fonction du nombre de combinaisons d'étiquettes possibles pour un segment. Au delà du seuil $\mathcal{S}$, la phrase est segmentée suivant la ponctuation (inter ou fin de phrase) ou en fonction des zones d'ambigüité faible. On espère ainsi découper le texte en segments pseudo indépendants et maximiser la probabilité globale d'appariement en maximisant celle de tous les segments.

Cependant, il arrive que le texte soit segmenté à des endroits sub-optimaux. Pour palier ce problème, nous avons ajouté à chaque segment les derniers mots du segment précédent après réaccentuation. Ces éléments sont bien sur supprimés avant la réunification finale.

4. Travaux liés

1. El-Bèze et al., 1994 [1]

Méthode également à base de désambiguisation morphosyntaxique. A la différence de la méthode présentée ici, ils utilisent des modèles de markov cachés d'ordre 3 (contre 2, ici) et une fenêtre de taille fixe. Ils obtiennent sur le corpus du journal "Le Monde" (années 91 et 92) un taux d'erreur de 0.36%.

2. Yarowsky, 1994 [2]

Méthode basée sur des listes de décisions qui combinent des étiquettes morphosyntaxique sur une fenêtre locale de 2 ou 4 mots et sur une fenêtre globale de 40 mots. Elle a été testée sur des données espagnoles mais, dans un second article [3], l'auteur a réalisé un comparatif de sa méthode avec une méthode très proche de celle présentée ici et a montré la supériorité de sa méthode.

5. Evaluation

Des tests réalisés empiriquement montrent que pour un seuil $\mathcal{S}$ supérieur à 16, le nombre d'erreurs converge asymptotiquement.

Notre évaluation a été réalisée sur plusieurs corpus désaccentués avec un seuil $\mathcal{S}$ fixé à 16. La ré-accentuation de ces corpus puis la comparaison avec le corpus original fournit les résultats présentés *figure 1*.

Nous pouvons constater que la grande majorité des erreurs sont dues à une mauvaise résolution morpho-syntaxique :

- Verbe au présent ou subjonctif (mange) ou verbe au participe passé (a mangé)
- Verbe (a) ou préposition (à)
- Conjonction de coordination (ou) ou adverbe (où)
- Article (la) ou adverbe (là).

Enfin, les mots inconnus accentués (absents de notre lexique) sont logiquement erronés puisque réécris tels quels.

Type of error	Number of occurrences	Percentage
-e VS. *-é* ending	155	32.6%
a VS. *à*	145	30.5%
ou VS. *où*	44	9.2%
la VS. *là*	33	6.9%
Unknown words	25	5.3%
Other	74	15.5%

FIG. 1 – Différentes erreurs commises par le système

Références

[1] El-Bèze, Merialdo, Rozeron, Derouault. "Accentuation automatique de textes par des méthodes probabilistes", Laboratoire Informatique d'Avignon, 97

[2] Yarowsky. "Decision lists for lexical ambiguity resolution : Application to Accent Restoration in Spanish and French", University of Pennsylvania, 94

[3] Yarowsky. "A Comparison of Corpus-based Techniques for Restoring Accents in Spanish and French Text", University of Pennsylvania, 94

[4] Merialdo. "Tagging English Text with a Probabilistic Model", Institut EURECOM, 1991.

[5] Baum. "An inequality and associated maximization technique. in statistical estimation for probabilistic functions of a Markov Processes" 1972.