Zodiac : Insertion automatique des signes diacritiques du français

Fabrizio Gotti et Guy Lapalme
RALI, Université de Montréal, CP 6128 Succursale Centre-Ville, Montréal, Canada, H3C 3J7
{gottif,lapalme}@iro.umontreal.ca

Résumé. Nous proposons dans cette démonstration de présenter le logiciel Zodiac, permettant l'insertion automatique de diacritiques (accents, cédilles, etc.) dans un texte français. Zodiac prend la forme d'un complément Microsoft Word sous Windows permettant des corrections automatiques du texte au cours de la frappe. Sous Linux et Mac OS X, il est implémenté comme un programme sur ligne de commande, se prêtant naturellement à lire ses entrées sur un « pipeline » et écrire ses sorties sur la sortie standard. Implémenté en UTF-8, il met en œuvre diverses librairies C++ utiles à certaines tâches du TAL, incluant la manipulation de modèles de langue statistiques.

Abstract. In this demo session, we propose to show how the software module Zodiac works. It allows the automatic insertion of diacritical marks (accents, cedillas, etc.) in text written in French. Zodiac is implemented as a Microsoft Word add-in under Windows, allowing automatic corrections as the user is typing. Under Linux and Mac OS X, it is implemented as a command-line utility, lending itself naturally to be used in a text-processing pipeline. Zodiac handles UTF-8, and showcases some useful C++ libraries for natural language processing, including statistical language modeling.

Mots-clés : aide à la rédaction, diacritiques, modèles de langue probabilistes.

Keywords: text editing, diacritical marks, statistical language models.

1 Contexte

Les *signes diacritiques* sont des symboles graphiques (accents, cédilles) combinés à des lettres déjà existantes afin d'en modifier la phonétique ou d'éviter la confusion entre des mots homographes (p.ex. « cote », « côte » et « côté »). Il en existe une vingtaine dans l'alphabet romain, et le français en utilise couramment cinq, soit la cédille, le tréma, les accents aigu, grave et circonflexe. Critiqués, on rencontre rarement le tilde (« cañon ») ou le rond en chef (« ångström »).

Malgré leur caractère indispensable à la sémantique du texte (« interne » ou « interné » dans un hôpital psychiatrique) et à la justesse orthographique, les diacritiques sont omis dans des situations relativement fréquentes. Les communications électroniques (courriels, textos, forums de discussion, etc.) présentent ces problèmes, que ce soit à cause de leur caractère informel, ou simplement parce que les méthodes de saisie rendent difficile l'insertion des diacritiques. Traditionnellement, ce dernier problème était surtout dû à des dispositions de clavier peu compatibles avec le français, mais l'avènement des tablettes et téléphones intelligents accentue désormais la difficulté à cause d'entraves ergonomiques. Par ailleurs, les apprenants de la langue française peuvent éprouver de la difficulté avec ces symboles.

Il peut donc s'avérer utile dans ces contextes de disposer d'un système d'insertion automatique de diacritiques à partir d'un texte qui en est dépourvu. La tâche n'est pas triviale, car un mot sans accents peut présenter plusieurs candidats avec diacritiques (voir l'exemple de « cote » plus haut). Il y a plus de dix ans, notre laboratoire proposait le logiciel Réacc, qui faisait l'insertion automatique des diacritiques (Simard & Deslauriers, 2001). Fondé sur un modèle de Markov caché, le logiciel offrait de bonnes performances, mais nécessitait le difficile maintien de librairies C faites sur mesure, dont l'instabilité a fini par condamner le logiciel. Une solution de rechange a donc été conçue.

Nous proposons ainsi une démonstration du logiciel Zodiac, un module effectuant la restauration de diacritiques automatiquement, et fondé sur un modèle statistique de langue et sur un lexique du français à large couverture. Notre démonstration montrera l'interface web de Zodiac (basée sur Linux), l'utilitaire ligne de commande, ainsi qu'un complément Zodiac à Microsoft Word permettant de faire l'insertion des diacritiques au cours de la frappe, de façon interactive, sous Windows.

2 Zodiac

2.1 Mode de fonctionnement

Zodiac commence par segmenter le texte entré en phrases et en mots, à l'aide de la librairie C++ ICU (`http://icu-project.org/`). ICU supprime ensuite tous les diacritiques déjà présents. Chaque jeton sans diacritiques est recherché dans une liste précompilée afin de trouver les candidats avec diacritiques (p.ex. : mais → mais ou maïs). Un modèle de langue statistique parcourt la phrase et utilise le contexte pour choisir une substitution candidate. Ainsi, « du maïs » sera préféré à « du mais ». Pour gérer l'explosion combinatoire lorsque des mots ambigus se suivent, une recherche en faisceau est utilisée. Le modèle de langue utilisé est un modèle trigramme entraîné avec la librairie SRILM (Stolcke, 2002). Le corpus d'entraînement consiste en 1 M de phrases des domaines politique et journalistique. Le modèle est compilé sous forme binaire avec la librairie C++ KenLM (Heafield, 2011) afin d'être chargé en 0,1 s en RAM.

2.2 Interface

La démonstration montrera les modules logiciels dans lesquels Zodiac est intégré. Une démo web est également disponible[1]. Le module est entièrement mis en œuvre en Unicode, grâce à la librairie ICU. **Sous Windows**, Zodiac est un complément Word à deux modes de fonctionnement : l'utilisateur peut activer l'insertion de diacritiques au fur et à mesure de la frappe ou il peut décider de sélectionner un passage d'un texte et d'en corriger les diacritiques avec Zodiac. **Sous Linux et Mac OS X**, la démonstration de Zodiac illustrera qu'il peut se comporter comme un filtre de texte typique, c'est-à-dire un programme lisant du texte UTF-8 et produisant une sortie textuelle corrigée. Il se prête donc à une utilisation par « pipeline », permettant le chaînage de plusieurs programmes.

2.3 Évaluation

L'évaluation d'un système d'insertion de diacritiques comme Zodiac est relativement simple. Il suffit de choisir un texte où les diacritiques figurent, et de le soumettre à Zodiac. Puisque celui-ci commence par supprimer les diacritiques pour ensuite les réinsérer selon l'algorithme décrit plus haut, les diacritiques produits en sortie sont exclusivement l'œuvre de Zodiac. L'évaluation consiste alors à comparer le texte original (la référence) avec la sortie de Zodiac (le candidat). Sur un corpus constitué de textes juridiques et littéraires, on constate que le système commet des erreurs en moyenne sur 0,54 % du nombre total de mots de corpus. La démonstration montrera certains aspects du logiciel affectant sa performance. Ainsi, des ambiguïtés subsistent pour l'insertion d'accents sur la première majuscule des noms propres (p.ex. « Eric » ou « Éric »). Cette difficulté est épineuse, à cause des variantes orthographiques et culturelles dans la rédaction des prénoms. Il reste également à déterminer s'il est constructif de supprimer les diacritiques préexistants avant leur insertion.

3 Perspectives

Une extension naturelle de Zodiac serait de s'intéresser à d'autres langues pour lesquelles des ambiguïtés comparables existent. Ainsi, le vietnamien utilise une orthographe appelée quốc ngữ, usant de nombreux diacritiques parfois combinés sur la même lettre, rendant la saisie malaisée. Zodiac utilise justement des librairies C++ capables de traiter du texte dans plusieurs langues, sans règles *ad hoc* spécifiques au français, donc cette voie nous paraît prometteuse.

Références

HEAFIELD K. (2011). KenLM : faster and smaller language model queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, p. 187–197, Edinburgh, Scotland, United Kingdom.

SIMARD M. & DESLAURIERS A. (2001). Real-time automatic insertion of accents in French text. *NLE*, **7**, 149–165.

STOLCKE A. (2002). SRILM - an extensible language modeling toolkit. In *In Proceedings Of The 7th International Conference On Spoken Language Processing (ICSLP 2002)*, p. 901–904.

[1] `http://rali.iro.umontreal.ca/rali/?q=fr/projet-zodiac`