Restoring accents in unknown biomedical words: application to the French MeSH thesaurus

Pierre Zweigenbaum, Natalia Grabar
Mission de Recherche en Sciences et Technologies de l’Information Médicale
STIM/DSI, Assistance Publique – Hôpitaux de Paris

STIM
CHU Pitié-Salpêtrière
91, boulevard de l’Hôpital
75634 Paris Cedex 13
France

{pz,ngr}@biomath.jussieu.fr
http://www.biomath.jussieu.fr/~{pz,ngr}/

1Initial versions of this work were presented at the EFMI Special Topic Conference Workshop on Natural Language Processing for Biomedical Applications [1], at the French Conference on Natural Language Processing [2] and at the ACL Workshop on Natural Language Processing in the Biomedical Domain [3]. The progression of this work owes to the comments of the reviewers and to the questions of the audience of these conferences.
Abstract

In languages with diacritic marks, such as French, there remain instances of textual or terminological resources that are available in electronic form without diacritic marks, which hinders their use in natural language interfaces. In a specialized domain such as medicine, it is often the case that some words are not found in the available electronic lexicons. The issue of accenting unknown words then arises: it is the theme of this work.

We propose two internal methods for accenting unknown words, which both learn on a reference set of accented words the contexts of occurrence of the various accented forms of a given letter. One method is adapted from part-of-speech tagging, the other is based on finite state transducers.

We show experimental results for letter $e$ on the French version of the Medical Subject Headings thesaurus. With the best training set, the tagging method obtains a precision-recall breakeven point of $84.2\pm4.4\%$ and the transducer method $83.8\pm4.5\%$ (with a baseline at 64%) for the unknown words that contain this letter. A consensus combination of both increases precision to $92.0\pm3.7\%$ with a recall of 75%. We perform an error analysis and discuss further steps that might help improve over the current performance.

Keywords

Natural Language Processing; Machine Learning; Controlled Vocabulary; Language; France; Algorithms