Toward Cleaner Web Corpora: Recognizing and Repairing Problems with Hybrid Online Documents

William H. Fletcher [1]

Abstract

Increasingly the Web is taken seriously as a source of documents for both ad-hoc and general-purpose corpora. In previous work I have explored techniques to improve the "signal-to-noise" ratio of texts retrieved from the Web by filtering out pages on the basis of document size and paragraph length. In two large corpora compiled recently from over 150,000 HTML webpages in English and Dutch, I found almost 5% of texts which passed these filters were too noisy for efficient and accurate POS tagging. The remaining noise was primarily due to "hybrid documents" with multiple or inaccurately declared character-set encodings (e.g. mixing several flavours of Windows, Macintosh or Unicode) or with islands of foreign language text and even digital data.

One approach is simply to identify and exclude hybrid texts–there are always more texts where they came from. However, since hybrids arise most frequently in multi-user environments with on-line discussion, doing so eliminates linguistically interesting spontaneous dialogue in favour of single-author monologue. Instead I attempt to salvage hybrid texts by recognizing and repairing inconsistent encodings and by bracketing out chunks in other languages.

Several challenges confront the would-be webpage salvager. The sequence and scope of character-set encoding reported at the server, document and sub-document levels must be observed strictly, but may be inaccurate. Since existing routines to verify and translate encoding assume consistency, they must be adapted to hybrid texts. Probabalistic methods based on the frequency of character n-grams are vital tools for language detection, but knowing a text's language is a prerequisite to character-set recognition.

This paper details the specific problems confronted and techniques developed in hopes of fomenting discussion and sharing solutions among researchers with similar challenges.

[1] United States Naval Academy
e-mail: fletcher@usna.edu