Automated detection of health websites' HONcode conformity: can N-gram tokenization replace stemming?

Célia Boyer[a], Ljiljana Dolamic[a] , Natalia Grabar[b]

[a] Health on The Net Foundation, Geneva, Switzerland, [b] Université Lille 3, France

Abstract

Authors evaluated supervised automatic classification algorithms for determination of health related web-page compliance with individual HONcode criteria of conduct using varying length character n-gram vectors to represent healthcare web page documents. The training/testing collection comprised web page fragments extracted by HONcode experts during the manual certification process. The authors compared automated classification performance of n-gram tokenization to the automated classification performance of document words and Porter-stemmed document words using a Naive Bayes classifier and DF (document frequency) dimensionality reduction metrics. The study attempted to determine whether the automated, language-independent approach might safely replace word-based classification. Using 5-grams as document features, authors also compared the baseline DF reduction function to Chi-square and Z-score dimensionality reductions. Overall study results indicate that n-gram tokenization provided a potentially viable alternative to document word stemming.

Keywords:

Machine learning; N-gram; HONcode.

Introduction

The HON Foundation's Code of Conduct[1], comprised of eight procedural guidelines, helps to indicate the credibility of online health information for both website editors and users. Currently, HON expert reviewers manually assess, re-assess, and certify health websites for compliance with the HONcode principles of conduct. The authors report herein a feasibility study to determine whether a specific machine learning algorithm based on n-gram representation of a document's content, can assist in the HONcode certification process.

Methods

In the scope of KHRESMOI[2] project we have developed a system for automated detection of HONcode principles [1]. Based on previous experience [1] the Naive Bayes machine learning algorithm is chosen in this study to compare the results obtained when various size n-grams (e.g. C3, C4, C5) and stems (W1p) are used as document tokens to the results of the word tokens (W1) baseline. The goal was to determine the extent to which words might be replaced as tokens while not sacrificing system classification performance. Keeping 30% of top ranked features, we also explored DF, Chi-square and Z-score dimensionality reduction for the 5-gram tokenization.

Results

Authors chose precision (P), recall (R) and F_1-measure to represent the quality of the classification. The results indicate that the tokenization method that produces the best precision results varies for each HONcode criterion. While W1 tokenization provides highest precision and highest F_1 value for "Authority" (0.64; 0.69) or "Privacy" (0.91; 0.94), highest precision for the "Reference" is obtained for 5-gram tokenization. For the "Justifiability" criterion the most balanced precision/recall trade off is achieved by C3.

The relative difference in precision between W1p and C5 spanning from -7.25% to 3.45% as well as similar behavior to W1 or W1p in relation with dimensionality reduction (e.g. features kept: 80% vs. 30%. Precision loss W1: -11,08%; W1p: -10.28%; C5: -10,41%) indicate the usability of C5 as an alternative to stem or word for the English language.

Results also show that both DF and Z-score dimensionality reduction significantly outperformed the Chi-square in terms of precision, with the exception of the "Contact details" criterion for the C5 tokenization.

Conclusion

Our study results indicate that the n-gram approach is a viable alternative to both word or stem tokenization. Choosing the "correct" dimensionality reduction algorithm can additionally improve the classification results. Accounting for importance of linguistic treat-

[1] http://www.healthonnet.org/HONcode/Conduct.html

[2] European project KHRESMOI (2010-2014, project No. 257528), http://khresmoi.eu

ment for morphologically complex languages, and the baseline established here, one can suppose that the language-independent n-gram approach would not only be interchangeable but also might result in better performance for languages other than English.

References

[1] Boyer C, Dolamic L. Feasibility of automated detection of honcode conformity for health related websites. IJACSA, 2014 Mar; 5(3):69-74

Address for correspondence

Célia Boyer, celia.boyer@healthonnet.org
Health on The Net Foundation
Chemin du Petit-Bel-Air 2
1225 Chêne-Bourg, Switzerland
+41 (0) 22 37 26 250