2017 Conference on Information Communications Technology and Society (ICTAS), Umhlanga, South Africa, 8-10 March 2017

Semi-supervised probabilistics approach for normalising informal short text messages

Modupe A
Celik T
Marivate, V
Diale, M

ABSTRACT:

The growing use of informal social text messages on Twitter is one of the known sources of big data. These type of messages are noisy and frequently rife with acronyms, slangs, grammatical errors and non-standard words causing grief for natural language processing (NLP) techniques. In this study, our contribution is to target non-standard words in the short text and propose a method to which the given word is likely to be transformed. Our method uses language model probability to characterise the relationship between formal and Informal-word, then employ the string similarity with a log-linear model to includes features for both word-level transformation and local context similarity. The weights of these features are trained by employing maximum likelihood framework using stochastic gradient descent (SGD) to hypothesise the better clean feature for a given informal short text. Experiments were conducted on a publicly available English-language tweet and the approach is able to normalise inflected words in an online social network.