South African Institute for Computer Scientists and Information Technologists (SAICSIT) 2013, 7-9 October 2013, East London, South Africa

A Comparison of Different Calculations for N-Gram Similarities in a Spelling Corrector for Mobile Instant Messaging Language

Laurie Butgereit[1,2]

[1]CSIR Meraka Institute Pretoria, RSA

[2]Nelson Mandela Metropolitan University Port Elizabeth, RSA

lbutgereit@meraka.org.za

Reinhardt A Botha[2]

[2]Nelson Mandela Metropolitan University PO Box 77000 Port Elizabeth, RSA

ReinhardtA.Botha@nmmu.ac.za

ABSTRACT

Mobile Instant Messaging (MIM) systems have produced a new convention in writing where vowels are often omitted, where new suffixes have appeared, where numerals and symbols often appear in the place of letters which have a similar shape or sound, and where words are often spelled phonetically. A word such as mister may be spelled numerous ways including mista and mistr (with new suffixes). When both participants to a MIM conversation understand these new spelling conventions, there is no problem. But in a situation such as automated topic spotting, it is advantageous to attempt to associate these new spellings (mista and mistr) back to the original word (mister). This paper describes work in creating a spelling corrector for MIM conversations for use after stop words have been removed from a conversation, after words have been stemmed, and after double letters have been collapsed to single letters. Four different similarity calculations Jaccard, Sørensen-Dice, Cosine, and Overlap are investigated and tested with historical data from the Dr Math mobile tutoring environment. This research found that the Overlap similarity calculation was the least accurate of the four measured. In situations where the length of the various words were the same, Sørensen-Dice and Cosine similarity calculations were identical. Jaccard and Sørensen-Dice worked equally well, however, they required different numerical cut-off values for misspelled words.