CNG Method with Weighted Voting

Vlado Kešelj and Nick Cercone ({`vlado,nick`}`@cs.dal.ca`)

CNG Method for Authorship Attribution. The Common N-Grams (CNG) classification method for authorship attribution (AATT) was described in [2]. The method is based on extracting the most frequent byte n-grams of size n from the training data. The n-grams are sorted by their normalized frequency, and the first L most-frequent n-grams define an author profile.

Given a test document, the test profile in produced in the same way, and then the distances between the test profile and the author profiles are calculated. The test document is classified using k-nearest neighbours method with $k = 1$, i.e., the test document is attributed to the author whose profile is closest to the test profile.

Given two profiles f_1 and f_2, which map n-grams from sets D_1 and D_2 to their respective frequencies, the distance measure between them is defined by the formula:

$$\sum_{g \in D_1 \cup D_2} \left(\frac{f_1(g) - f_2(g)}{\frac{f_1(g)+f_2(g)}{2}} \right)^2 = \sum_{g \in D_1 \cup D_2} \left(\frac{2 \cdot (f_1(g) - f_2(g))}{f_1(g) + f_2(g)} \right)^2 \tag{1}$$

where $f_i(g) = 0$ if $g \notin D_i$.

The CNG method achieved state-of-the-art performance in AATT for English, Greek, and Chinese [2]. The method relies on two parameters, n and L, so a remaining issue is how to avoid the parameter dependency.

CNG Method with Weighted Voting (CNG-wv) — Run 1. In the previous experiments on English and Greek texts, it was noted that the CNG method achieved high performance for parameter values $3 \leq n \leq 8$ and $1000 \leq L$. For practical reasons we introduce the limit $L \leq 5000$.

For each combination of parameter values n and L, $(n, L) \in \{3, 4, 5, 6, 7, 8\} \times \{1000, 2000, 3000, 4000, 5000\}$, we calculate a vote for a best-matching label, and by summarizing and sorting the votes we obtain the final ranking. The weight of a vote is calculated in the following way: Let a be the similarity of the best-matching label A, and b be the similarity of the second best. The ratio $r = 1 - a/b$ gives the weight of the vote for the label A. In order to obtain probability distribution, the vote sums are normalized.

CNG-wv Method with reject — Run 2. The CNG method classifies texts into given training classes. However, it is not defined how to reject all the training classes, i.e., how to detect that the author of the text is not among authors whose training samples are provided. We have adjusted the method CNG-wv for this purpose in the following way: After a probability distribution is obtained for the given author labels, if the best-matching author label has probability $p < 0.5$, we introduce the best-matching label 'OTHER,' meaning reject, with the 'probability' $1 - p$, and we re-normalize probabilities.

The Perl module Ngrams [1] is used to obtain the profile of the n-grams of type *byte.*

References

[1] Vlado Kešelj. Perl package Text::Ngrams, 2003. Accessed in May 2004.
`http://www.cs.dal.ca/~vlado/srcperl/Ngrams` or
`http://search.cpan.org/author/VLADO/Text-Ngrams-1.1/`.

[2] Vlado Kešelj, Fuchun Peng, Nick Cercone, and Calvin Thomas. N-gram-based author profiles for authorship attribution. In *Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING'03*, pages 255–264, Dalhousie University, Halifax, Nova Scotia, Canada, August 2003.