Authorship Detection

Minutes for 6th meeting on 9th Sep 2010
Held in Level 3, meeting room, Innova21, 1.30PM – 2.30PM
Participants: Clement, Joel, Jie Dong, Brian Ng
Absent: None

Minutes

1. **Project status/progress during last week**
   a) Simple Trigram Markov model has been developed in JAVA by Jie
   b) Code of Word recurrence interval method is developed by Clement, it is able to calculate the standard deviation of WRI of the selected key function word in the text
   c) Part of Word frequency method code is written by Joel
   d) Present our work to Brian and ask for feedback

2. **Project goals for upcoming week**
   a) Code implementation on three statistical text extraction algorithms
      i. Joel: Word Frequency Interval
      // Things need to be modified
      ii. Clement: Word Recurrence Interval
      The algorithm produces too much data. It is needed to revise and cut down on the data by reducing the number of key function words to be used.
      For SVM, it is advisable to have standard results after the data extraction process. Normalise the length of the vectors. Furthermore, it is also advised to combine different kind of approach to produce interesting results.

      iii. Dong Jie: Trigram Markov:
      The previous algorithm only considers effect of the trigram words. Result for a test paragraph contains a lot useless information, which about 70% of trigrams only appear once. Information which is worth using in classification is just about 10%. By extracting common trigrams from several test texts, few of them left. Hence, another enhanced model, in which unigrams and bigrams are also taken into consideration, will be tested in the following week.
      In addition, SVM will also be used to test the result in coming week. Investigating how to use SVM functions in MATLAB, svmtrain and svmclassify (Bioinformatics toolbox)

   b) Peer review on Stage 1 design document – Audio Assisted Vision System