Classifying news stories using memory based reasoning

Method: Memory based reasoning.

350 different codes

The simplest formulation: Solves a new task by looking up examples of tasks similar to the new task and using similarity with these remembered solutions to determine the new solution.

\*Codes are assigned to new unseen stories by finding new matches from the training database and then choosing the best few codes based on a confidence threshold.

Classification algorithm (general approach of MBR)

1. Find the near matches for each document to be classified.
2. Assign codes to the unknown document by combining the codes assigned to the *K* nearest matches (up to 11 nearest neighbors)
3. Codes are assigned weights by summing similarity scores from the near matches.
4. We choose the best codes based on a score threshold.

In the project: single words and capital words pairs were used as features.

SEEKER: the match engine

1. Eliminates stop words (368 non-content bearing words, like “the”, “on”)

Eliminates the most common words that account for 20% of the database.

1. Removes 72 additional words (? There were no examples)
2. Remaining words (searchable terms) are assigned weights inversely proportional to their frequencies in the database.

There were over 250.000 searchable words and word pairs in this database.

Variation of different parameters

* Varying the confidence threshold: variation of recall and precision.
* Using different number of near matches: As the number of near matches increases more correct codes are found but also more noise is added. Optimal combination requires further study.
* Weight of capital pairs: This increase or decrease classification relevance based on proper names.

Evaluation

N-way cross validation: excluding each test example one at a time from the database and perform the classification on it. 1000 articles were used for the test set.

Performance reported here is the average for the entire database.