Evidence Grade Classification

Timothy Baldwin (and a cast of 57)

Dept of Computer Science and Software Engineering
The University of Melbourne

Overview

Task: Given a set of documents, determine the "evidence grade" (A, B or C)

Approach: Farm out to 57 3rd-year CS students as part of the assessment of the *Knowledge Technologies* subject

Finding: Some highly successful methods proposed ... if only the students had submitted them formally!

A Sample of Approaches

Basic Approach

1. Map set of abstracts into single meta-document, and further map the meta-document into features
2. Train a supervised model off the training instances
3. Apply the learned model to the development/test documents

Variants on a Theme

- word/stem features (possibly indexed based on source, e.g. title vs. journal vs. abstract)
- metadata
- no. documents returned
- feature weighting/selection
- different learners (k-NN, NB, NP, random forest, SVM, ...)
- meta-classification over the systems from a single student
- approach problem as constraint satisfaction problem, interpreting SORT code for given query as upper bound for SORT code for individual document

Mega Meta-classifier

- Also played around with stacking-based meta-classification, based on all the student systems trained over the training data, and applied to both the dev and test data
- Because of inconsistencies in submissions, only 23/91 systems could actually be used for meta-classification
- As the meta-learner, used a support vector regression model, mapping ordinal categories onto fixed-interval real values, and discretising the results back to the ordinal categories

Basic Findings

- Good feature representation with feature selection tends to do best; little gain from metadata
- The choice of learner had relatively little impact on results
- Students found the task much harder than Sarker et al. (2011) suggested
- Mega meta-classifier disappointing, partly due to alignment issues
- Slightly opaque nature of queries/annotation process was slightly confusing
- Great task to get hands-on experience with language technology/machine learning (dressed up as "Knowledge Technologies")

Results

Methodology	Dev F-score	Test F-score
Majority class baseline	0.45	0.49
k-NN + stemmed words + meta-data	0.53	0.54
SVM meta-classifier (words, meta-data, etc.)	0.55	0.50
SVM + words + feature selection	0.62	0.52
Constraint satisfaction	0.41	0.28
⋮	⋮	⋮
Mega meta-classifier	—	0.49