HRCS Auto-Coding using Elsevier Technology

Giles Radford, Custom Solution Manager

22 November 2011

Background

- Elsevier has been using the ‘Fingerprint’ technology to assist funding agencies around the world with software that assists with peer reviewer selection, reporting, program planning etc.
- Fingerprinting can classify applications etc. against broad thesauri such as MeSH, Compendex, Geotree etc. but NOT HRCS, CSO, ICD10 etc.
- By utilizing a combination of the Fingerprint technology with another technique called ‘Support Vector Machine’ we have assisted CRUK with their CSO coding and will implemented auto-coding for HRCS HC and RAC, and ICD10, for NETSCC. (NIHR Evaluation, Trials and Studies Coordinating Centre)
- SVM uses a ‘learning set’ (previously coded grants) to create an algorithm that can replicate coding and then apply it going forward.

Process

- All the applications are fingerprinted against an appropriate thesaurus – MeSH for HRCS and MeSH/NCI for CSO coding
- All existing codes are examined and an algorithm is created that tries to emulate manual coding
- Outliers are examined (manually) and the algorithm adjusted to attempt a better match to learning set. Repeat.
- System is installed in host organization's system. In both CRUK and NETSCC we are installing a ‘suggestion’ system with final manual over-ride possible. Changes are fed back into the learning set for continuous improvement.
- Fully-automatic is possible.

Findings

- Analysis at CRUK suggests:-
- About 50% taken as is
- About 75% taken with minor adjustment
- About 90% are ‘acceptably’ coded

- This is year 1 of algorithm. Feedback of these results will improve figures next year.

Issues

- Difference in automated and actual may have many causes:-

1. Limited dataset (especially on certain terms) reduces ability to predict
2. Inconsistent manual allocation creates inconsistencies in the vectors: same data in, but different data out, naturally confuses the algorithm
3. Replicating HRCS ‘rules’ is a problem. Next slide

HRCS Rules

- HRCS coding rules:-
  - RAC (‘capture the main objective of the research’)
  - Use a maximum of **two** codes unless coding a large programme of research, in which case up to 4 codes can be used
  - HC (‘captures the area of health or disease being studied’).
  - A maximum of **five** categories can be applied if a number of different areas of health or disease are included in the study. These should be equally apportioned unless clearly stated otherwise in the abstract

Issues for fully automated system

- 75% of HRCS RAC codes had a single entry, so:-
- We could replicate that finding, and apply a single term in 75% of cases, OR
- We use a percentage rule. So, if two terms are returned 50%:50% then we apply two terms, and if two terms are returned 90:10 then we apply one
- Best approach is to combine both: return 75% with single code and make that split by examining percentage allocation to terms returned

Conclusions

- We believe we can automated HRCS coding for both RAC and HC to a degree of accuracy that is (almost) as good as manual coding
- We can auto-code entire back–history
- We are working with Ian Viney to consider how best to install such a system, considering the process, the system, interfaces etc.

THANK YOU

Name g.radford@elsevier.com