HGVS Scientific & Annual General Meeting
November 4, 2003
Los Angeles CA, U.S.A.

Extraction of cancer-specific genomic variation from the biomedical literature

Winters RS[1], McDonald RT[2], Jin Y[1], Kim J[1], Wooster R[5], Liberman MY[2,3], Pereira F[2], White PS[1,4]

[1]Division of Oncology, Children's Hospital of Philadelphia; Departments of [2]Computer and Information Science, [3]Linguistics, and [4]Pediatrics, University of Pennsylvania, [5]The Wellcome Trust Sanger Institute

Consolidation of the totality of genomic variation events identified in human malignancies would be an invaluable information resource for the oncology research community. Currently, much of this information is disparately localized within biomedical research articles. Accordingly, we have developed a strategy to mine the biomedical literature for human variation events described in the biomedical literature using a comprehensive information extraction approach. A set of 650 PubMed full-text articles encompassing eight genes commonly mutated in cancer was identified and manually annotated for 28 entity classes associated with malignancy and genomic variation. This data was used as a validation set for our procedure, and abstracts from the set were used as training material for manual and automated annotation procedures. Three entity classes were targeted: *genes*, *variation events*, and *malignancies*. After annotation of a pilot set of abstracts, the entity class *variation events* was separated into component classes *variation type*, *variation location*, *initial genomic state*, and *subsequent genomic state*. Strict definitions of each class were authored and adhered to during annotation. Initially, the abstracts underwent manual annotation in three areas: part-of-speech annotation, entity tagging, and predicate-argument (syntactic) analysis. A fourth component, reference resolution and relation tagging (semantic analysis), is in progress. Abstracts are annotated once completely and then checked in a 2nd pass by domain expert annotators; a small percentage of abstracts are independently dual-annotated to completion. Results are then adjudicated by a senior domain expert, and conflicts are discussed and resolved collectively. A series of annotator-assistive computational tools with graphical interfaces are used for annotation work. Annotation results are stored in a separate, associated XML document. Workflow infrastructure and relational database systems were implemented for streamlining process and result management. Manually annotated abstracts were then used to design automated entity tagging algorithms for the *gene* and *variation* subcomponent entity classes. An automated tagging algorithm capable of recognizing the variation subcomponents *type*, *location*, and *state* was constructed using the MALLET toolkit (http:// http://www.cs.umas#s.edu/~mccallum/mallet/). The manually annotated files were used to train a conditional random field model. A series of expertly defined rules, based upon literature observations, augments the basic algorithm (i.e. "[{1-22, X, Y}{p,q}] may be associated with a location type"). Our initial evaluation was performed on 50 abstracts annotated by a senior domain expert and used a stringent evaluation criteria requiring both the predicted tag and its exact boundaries to be correct. Initial results, with 10-fold cross validation, yielded an f-score of 0.69 with high precision (0.84) but relatively low recall (0.58). Presently, we are increasing the size of the training corpus, anticipating an increase in recall. It is expected that our annotation pipeline will continue to be perfected and more fully automated via an iterative cycle of annotation exercises, analysis of results, and procedure/algorithm modification. Over time, the results of this process, when extended to full-text articles, can be combined with existing cancer variation datasets to provide a more comprehensive set fully integrated with the biomedical literature.