Generalizability of Hybrid Search Algorithms to Map Multiple Biomedical Vocabulary Domains

Senthil K. Nachimuthu, MD[a,b] and R. Dean Woolstenhulme[a]
a. 3M Health Information Systems, Salt Lake City, Utah.
b. Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah.

Abstract: *Hybrid text matching algorithms similar to those used for DNA sequencing were developed by 3M Health Information Systems to map a noisy legacy codeset to the 3M Healthcare Data Dictionary (3M HDD). Applying these techniques to map other biomedical vocabularies was briefly introduced in an earlier paper describing the algorithms [1]. We now present results from successfully utilizing them to map different vocabularies across multiple biomedical domains, proving their generalizability.*

INTRODUCTION

Multiple biomedical vocabularies are often used by clinical information systems. Interoperability between these vocabularies necessitates mapping them to each other, or to a Reference Terminology such as the UMLS Metathesaurus® or the 3M HDD. Mapping a concept may require multiple searches to find a match, involving searching for synonyms, abbreviations, expansions, word or spelling variants, etc. This can be a challenging task, and is wasteful of the domain expert's time, which can be better spent understanding and mapping the concepts rather than manually searching for such variations.

BACKGROUND AND METHODS

3M Health Information Systems was required to map a locally developed legacy codeset which had highly inconsistent surface forms. This necessitated the development of search tools that could automatically search for all possible permutations of the input term. We developed a hybrid search algorithm (named HyperSearch), which searched the input term against the 3M HDD by using synonyms, abbreviations, abbreviation expansions, word stemming and ignoring word order, typos and stop words[1]. The UMLS SPECIALIST Lexicon and a custom knowledgebase for the legacy codeset were highly utilized in this process. HyperSearch mapped more than 70% of the legacy codeset, compared to less than 20% with SQL searches alone.

Hypersearch produced the results as a list sorted by match score. It was designed and built as modules which can be can be integrated as required. Removing the custom knowledgebase module from HyperSearch allowed it to be used for mapping any biomedical vocabulary, instead of just the legacy codeset it was developed for.

RESULTS

Results from mapping a Lab Codeset (Table 1) and a subset of SNOMED CT® (Table 2) to the 3M HDD using HyperSearch and SQL queries are presented below. A positive match is when the exact match is within the top 25 results returned. *Total Matched* column denotes the number of concepts that were matched to the HDD by all methods. *HyperSearch* and *SQL Search* denote the number of matches (out of *Total Matched*) that were returned by HyperSearch and SQL Queries respectively. *SCT – LOINC* column denotes the number of matches found by using the SNOMED CT – LOINC Integration Table provided by SNOMED.

LAB	Total Matched	Hyper Search	SQL Search	SCT - LOINC
All	5,920	5,351	339	3,406
Analyte	4,689	4,689	258	3,078
Method	290	290	NA	172
Specimen	243	243	82	80

Table 1. Mapping a Lab Codeset to the 3M HDD

SNOMED CT®	Total Matched	Hyper Search	SQL Search
Concepts	69,148	68,837	14,990

Table 2. Mapping SNOMED CT® to the 3M HDD

These results show that HyperSearch found almost all the matches that were present in the 3M HDD, including those missed by SQL Searches. This study proves the generalizability of our search algorithms, and the value of advanced search algorithms for biomedical vocabulary mapping.

References

1. Nachimuthu SK. and Lau LM. Applying hybrid algorithms for text matching to automated biomedical vocabulary mapping. AMIA Annu Symp Proc 2005.