ASSOCIATE PROJECT PROPOSALS

Project Title:
Enhanced Synonym Mapping Tool (SMT) for Concept Mapping by Integrating Lexical Tool Fruitful Variants

Submitted By: Allen Brown

Brief Description (approximately 500 words)

The Sub-Term Mapping Tools (STMT), developed by Lexical Systems Group (LSG), is distributed by NLM via an Open Source License agreement. STMT is a generic tool set, with fully configurable options (corpus, synonyms, etc.), that provides comprehensive sub-term related features for query expansion and other NLP applications with Java APIs and command line tools. The Synonym Mapping Tool (SMT) is one of the most commonly used tools in the STMT package and is designed to find concepts in the UMLS-Metathesaurus using synonym substitutions. First, it loads the corpus of normalized synonyms in a tree structure with each term as a branch in the tree and each word in the term as a node in the branch. Second, sub-terms of the input term that have synonyms are found by traversing through the corpus tree from the root node [1]. For example, "decubitus ulcer", "ulcer", and "area" are found as sub-terms of "decubitus ulcer of sacral area" from this algorithm. However, "of" and "sacral" are not sub-terms because they are not in the corpus of synonyms. Third, sub-term patterns with less than two sub-terms are found: 1) patterns with one sub-term: "decubitus ulcer of sacral area", "decubitus ulcer of sacral area", and "decubitus ulcer of sacral area"; and 2) patterns with two sub-terms: "decubitus ulcer of sacral area" and "decubitus ulcer of sacral area". Finally, synonyms are substituted for sub-terms in each pattern to form new terms for concept mapping. In this example, "Pressure ulcer of sacral region" with C2888342 is found for the term with two sub-term synonyms substituted for "pressure ulcer of sacral region", where "pressure ulcer" and "region" are synonyms of sub-terms "decubitus ulcer" and "area", respectively. By applying this query expansion technique in UMLS-CORE project [2], SMT is able to increase about 10% of the coverage rate on CUI mapping with the same accuracy.

The performance (coverage and accuracy) of SMT in finding the mapped concepts mainly depends on the synonym corpus if the number limit of substituted synonyms is fixed. Terms have same, similar, or related meanings are considered as synonymous terms in SMT for the sub-term substitution to improve the coverage rate without dropping the accuracy. Accordingly, lexical variants, such as spelling variants, inflectional variants, acronym/abbreviations, expansion, and derivational variants, may be used. The fruit variants in Lexical Tools package combines these lexical variants with a distance score [3], which are shown in table-1.

Operation	Notation	Distance score
Spelling Variant	s	0
Inflectional Variant	i	1
Acronym/Abbreviation	A	2
Expansion	a	2
Derivational Variant	d	3

Table-1 Fruitful variants operations with distance score

It would be beneficial to NLP research if we could 1) identify the weight of each lexical variant contributes to the coverage rate; 2) find the combination of these lexical variants for optimum performance in query expansion. This study can be accomplished by applying different lexical variants corpora to SMT for concept mapping as follows. First, a list of terms should be collected as a testing set. Second, apply MetaMap [4] with experts' review to find the correct concept of all terms in this testing set as a gold standard. Third, compare the recall and precision rate of SMT with various synonym corpora to the gold standard. Finally, construct a table of optimum fruitful variants for all Lexical terms and integrate it in SMT for best CUI mapping.

Main Research Steps:

The following section briefly describes the high level procedures for this study:

I. Define the testing set and gold standard
   - Use the testing set from UMLS-Core or MetaMap project
   - Or conduct a new term list and use as testing set
   - Use concept found from MetaMap and with help from experts' review as gold standard

II. Result: Compare the SMT with various corpora
   - Find all lexical variants for terms in Lexicon
   - Construct SMT synonyms corpus from the found lexical variants
   - Find concepts from original SMT and enhanced SMT with various corpora
   - Compare the system performance, precision and recall for above methods to find optimized fruit variants
   - Integrate the found fruit variants table with SMT for best CUI mapping

References:

1. Lu, C. J. Lu and, A. C. Browne, "Development of Sub-Term Mapping Tools (STMT)", AMIA 2012 Annual Symposium, Chicago, IL, November 3-7, 2012, p. 1845

2. K.W. Fung, C. McDonald, S. Srinivasan, "The UMLS-CORE project: a study of the problem list terminologies used in large healthcare institutions", JAMIA, 2010, Vol. 17, p.675-680.
3. A.R. Aronson, "The Effect of Textual Variation on Concept Based Information Retrieval". In Cimino J(ed) *Proceedings of the AMIA Fall Symposium*, 373-377, 1996
4. A.R. Aronson and F.M. Lang, "An Overview of MetaMap: historical perspective and recent advances", JAMIA, 2010, Vol. 17, p.229-236

Total Duration / Elapsed Time [in weeks]: 12 ~ 24

External Schedules / Deadlines [if any]:

Primary Learning Objectives for Associate:

- Learning Natural Language Processing
- Learning fruitful variants
- Learning how to use MetaMap
- Learning STMT and Query expansion
- Learning life-cycle software development
- Learning Java, Html, shell scripts
- Paper publication

Expected Project Experiences (select from the list):

Problem definition
Project scope definition
Design and implementation of research methodology
Use of applied statistics
Data analysis
~~Workflow analysis~~
Development of functional specifications
Identification of and negotiation for needed project resources
Examination of an unfamiliar technical area
Identification of others' technical expertise
Identification and evaluation of alternatives
Development and presentation of recommendations
~~Responsibility and accountability for a discrete product~~
~~Role definition in a task group and participation in group dynamics~~
~~Observation of supervisory activities (e.g., personnel assignment, training, development of procedural guidelines)~~
~~Observation of management styles~~
~~Observation of organizational politics~~
Preparation of a manuscript for publication

Expected Outputs/Products:

- Technical reports or paper publication

Suggested Methodologies (select from the list):

Associate will develop the methodologies in conjunction with the project leader
Literature Review
Websites reviews
~~Environmental Scan~~
~~Interviews~~
Data analysis
~~Focus Groups~~
~~Usability testing~~
Web site creation
~~Distance education materials – creation~~
Other:

Benefits to NLM:

- Evaluate the powerfulness of fruitful variants in query expansion technique
- Establish a standard testing set for concept mapping in NLP
- Paper publication

Project Leader(s):

- Allen Brown

Other Resource People:

- Chris Lu

Software/server access required:

- Linux
- Window
- Java
- Ant
- Html
- Shell scripts
- Apache Httpd server