Development of Sub-Term Mapping Tools (STMT)

Chris J. Lu, Ph.D.[1, 2] and Allen C. Browne [1]

[1]National Library of Medicine, Bethesda, MD; [2]Medical Science & Computing, Inc, Rockville, MD

Abstract

The Sub-Term Mapping Tools (STMT), developed at National Library of Medicine (NLM), are used to find 1) all sub-terms in a specified corpus; 2) the longest prefix sub-term; 3) all sub-term patterns; and 4) concept mapping of all permutations on synonym substitutions for a term. It has been successfully used in MetaMap and UMLS-CORE projects. STMT is distributed by NLM via an Open Source License agreement.

1. Introduction

A sub-term is a term that is a subset of another term. In general, a term is composed of words (delimitated by space/tab). Finding sub-terms, in a specified corpus, is an important application in Natural Language Processing (NLP) projects. However, no generic tool is available. For example, MetaMap identifies sub-terms (known to Metathesaurus) as candidates and then evaluates them with scores to find the best concept mapping [1]. One of the key steps is to find the longest prefix sub-term on an input term that is a lexical item (LexItem) in the SPECIALIST Lexicon. The longest prefix usually leads to the best concept match. The UMLS-CORE project substitutes sub-terms with synonyms for those terms that do not have normalized string matches in Metathesaurus [2]. The LexItem Sub-term Finder (LSF) and Synonym Mapping Tool (SMT) are used in above projects via STMT APIs with pre-load corpora. In the following we describe the features of STMP.

2. LexItem Sub-term Finder (LSF)

LSF is used to find the longest prefix and sub-terms that are known to Lexicon. LSF uses a pre-assigned corpus including normalized inflectional variants from the Lexicon. It loads the corpus to the STMT 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. The longest prefix LexItem and sub-term LexItems can be retrieved quickly by matching each word along the input term to nodes on the traversing path in the corpus tree. This fast algorithm replaces numerous expensive LexAccess calls to improve performance in MetaMap. Let's assume the input term is "decubitus ulcer of sacral area". "decubitus ulcer" is found as the longest prefix LexItem (sub-terms are in underlined) and "decubitus", "decubitus ulcer", "ulcer", "of", "sacral", and "area" are found as sub-terms known to Lexicon.

3. Synonym Mapping Tool (SMT)

SMT is used in the UMLS-CORE project to find concepts for synonym substitutions. First, it uses a pre-assigned corpus including all normalized terms that have synonyms. Second, sub-terms of the input term are found. In this case, "decubitus ulcer", "ulcer", and "area" are considered sub-terms of "decubitus ulcer of sacral area". But "of" and "sacral" are not sub-terms because they are not in the corpus of synonyms. Third, sub-term patterns with specified numbers of 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, C2888342 is found for the term with two sub-term synonyms substituted "pressure ulcer of sacral region", where "pressure ulcer" and "region" are synonyms of sub-terms "decubitus ulcer" and "area", respectively.

4. Conclusion

STMT is generic tool set with fully configurable options (corpus, synonyms, etc.) that provides comprehensive sub-term related features for NLP applications with Java APIs and command line tools. This tool is distributed with Lexical Tools and freely available at: http://specialist.nlm.nih.gov/lvg.

References

1. A.R. Aronson and F.M. Lang, "An Overview of MetaMap: historical perspective and recent advances", JAMIA, 2010, Vol. 17, p.229-236
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