Scalability of Piecewise Synonym Identification in Integration of SNOMED into the UMLS

Kuo-chuan Huang, New Jersey Institute of Technology, USA
James Geller, New Jersey Institute of Technology, USA
Michael Halper, New Jersey Institute of Technology, USA
Gai Elhanan, New Jersey Institute of Technology, USA
Yehoshua Perl, New Jersey Institute of Technology, USA

ABSTRACT

Synonym identification during source terminology integration into the Unified Medical Language System (UMLS) is a labor-intensive task needed for every new release of the source. The piecewise synonym (PWS) methodology was previously used for the integration of a small source. The goal of this paper is to determine whether the piecewise synonym methodology with two control parameters scales to a much larger terminology (a subset of SNOMED CT), the control parameters are necessary to make the methodology viable, and the control parameters lead to any loss of matching results. Additional methods for limiting the size of the dictionary used in the PWS generation methodology are used. The authors’ methodology discovered 41% of concepts not found by string matching. The necessity and effectiveness of the control parameters were confirmed. Furthermore, when comparing the results of experiments with and without control parameters, no matches were lost.

Keywords: SNOMED CT, Source Integration, Synonym Generation, Synonym Identification, Synonym Substitution, Synonymy, Term Decomposition, Term Matching, Unified Medical Language System (UMLS)

INTRODUCTION

The integration of biomedical terminologies is a labor-intensive, error-prone task. The use of string matching methods alone is not sufficient to solve this problem. One environment where the terminology integration task has to be performed on a massive scale is with every new release of the Unified Medical Language System (UMLS). The UMLS (Bodenreider,
2004; Humphreys, Lindberg, Schoolman, & Barnett, 1998; Lindberg, Humphreys, & McCray, 1993) is a large terminological database containing biomedical terms from many source terminologies. At every new release of the UMLS (currently biannually) new source terminologies are integrated. Updates or new versions of old source terminologies are also reintegrated into the UMLS. For example, the Gene Ontology (GO) (Gene Ontology Consortium, 2010), originally integrated into the UMLS in 2004 (Lomax & McCray, 2004), had about 25,000 concepts in the UMLS version 2008AA. However, the number of GO concepts increased to more than 48,000 concepts in 2008AB.

The goals of the UMLS are to overcome two problems, the distribution of useful biomedical information among disparate databases and systems and the variety of ways the same concept is expressed in different sources. The UMLS contains terminologies from different medical domains, forming a large terminological repository to solve those two problems (Cimino, 1998; Humphreys et al., 1998). However, since the repository is large, the fact that “the same concept may be expressed in many different ways in different sources” (Humphreys et al., 1998) becomes a difficulty when integrating a new source terminology into the UMLS. It is sometimes difficult to match a term from a new source with the correct concept in the UMLS, even with the help of lexical tools provided by the National Library of Medicine, such as MetaMap and Norm (Cantor et al., 2003). Thus, one major problem during UMLS source integration is the identification of terms and associated concepts from the new source that already exist in the UMLS.

In the UMLS 2008 AB version, the terminological repository, called Metathesaurus (Schuyler, Hole, Tuttle, & Sherertz, 1993; Tuttle et al., 1990), contains 147 source terminologies with more than 2 million concepts and over 9 million terms (U. S. National Library of Medicine, 2010c). Among these sources, SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms) (IHTSO, 2010) may be considered one of the most important, due to two factors: The number of the concepts in SNOMED CT (July 2008 version) is roughly 380k, which is the largest (English) UMLS source measured in terms. Contrary to many other UMLS sources, SNOMED CT has a rich structure, which is based on a formal model, namely a version of Description Logic (Campbell, Das, & Musen, 1994; Spackman, 2001). For a report on the original integration of SNOMED CT into the UMLS see Fung et al. (2005). Like the UMLS, SNOMED CT is updated twice a year, and changes need to be migrated into new releases of the UMLS.

In this paper, we are continuing our study of the use of two non-syntactic techniques for finding new synonyms for given multi-word terms (Huang, Geller, Halper, & Cimino, 2007; Huang, Geller, Halper, Perl, & Xu, 2009), namely extraction and substitution, together defining the piecewise synonym (PWS) methodology, but we are focusing on their scalability. Extraction and substitution are used together with string matching. Below follows an informal explanation of extraction and substitution. A precise description is given in the Background Section.

In the extraction (preprocessing) stage, new synonyms are generated from existing multiword UMLS synonyms. The result of this preprocessing stage is a dictionary of synonyms, which we call Generalized Synonym Dictionary (Huang et al., 2007). For example, in the UMLS the terms “Artificial lens” and “Prosthetic lens” are synonyms. Extraction (preprocessing) eliminates the common word “lens” and postulates that “Artificial” and “Prosthetic” are synonyms. This fact is stored in the Generalized Synonym Dictionary.

In the substitution stage, a term from the new source is decomposed into its constituent words and all applicable synonyms (from the UMLS and the Generalized Synonym Dictionary) are retrieved for each word individually. New terms are then created by substituting these synonyms for the original single words in the term, while preserving the order. Note that an original word may remain at its position in the substitution process, but not all of them. For example, with the synonyms (Artificial, Pros-