Artificial Intelligence in Medicine (2005) **34**, 219–233

ELSEVIER

ARTIFICIAL INTELLIGENCE IN MEDICINE

http://www.intl.elsevierhealth.com/journals/aiim

An expert study evaluating the UMLS lexical metaschema

Li Zhang [a], George Hripcsak [b], Yehoshua Perl [a,*], Michael Halper [c], James Geller [a]

[a] *Computer Science Department, New Jersey Institute of Technology, University Heights, Newark, NJ 07102, USA*
[b] *Department of Biomedical Informatics, Columbia University, New York, NY 10032, USA*
[c] *Mathematics and Computer Science Department, Kean University, Union, NJ 07083, USA*

Received 24 June 2004; received in revised form 10 January 2005; accepted 10 January 2005

KEYWORDS
Lexical partition;
Lexical metaschema;
UMLS;
Semantic network;
Expert study;
Evaluation

Summary

Objective: A metaschema is an abstraction network of the UMLS's semantic network (SN) obtained from a connected partition of its collection of semantic types. A *lexical metaschema* was previously derived based on a lexical partition which partitioned the SN into semantic-type groups using identical word-usage among the names of semantic types and the definitions of their respective children. In this paper, a statistical analysis methodology is presented to evaluate the lexical metaschema based on a study involving a group of established UMLS experts.
Methods: In the study, each expert was asked to identify subject areas of the SN based on his or her understanding of the various semantic types. For this purpose, the expert scans the SN hierarchy top-down, identifying semantic types, which are important and different enough from their parent semantic types, as roots of their groups. From the response of each expert, an ''expert metaschema'' is constructed. The different experts' metaschemas can vary widely. So, additional metaschemas are obtained from aggregations of the experts' responses. Of special interest is the *consensus metaschema* which represents an aggregation of a simple majority of the experts' responses. Statistical analysis comparing the lexical metaschema with the experts' metaschemas and the consensus metaschema is presented.
Results: The analysis results shows that 17 out of the 21 meta-semantic types in the lexical metaschema also appear in the consensus metaschema (about 81%). There are 107 semantic types (about 79%) covered by identical meta-semantic types and refinements. The results show the high similarity between the two metaschemas.

* Corresponding author. Tel.: +1 973 596 3392; fax: +1 973 642 7029.
E-mail addresses: lxz1853@njit.edu (L. Zhang), hripcsak@columbia.edu (G. Hripcsak), yehoshua.perl@njit.edu (Y. Perl), mhalper@kean.edu (M. Halper), james.geller@njit.edu (J. Geller).

0933-3657/$ – see front matter
doi:10.1016/j.artmed.2005.01.002

Furthermore, the statistical analysis shows that the lexical metaschema did not grossly underperform compared to the experts.
Conclusion: Our study shows that the lexical metaschema provides a good approximation for a partition of meaningful subject areas in the SN, when compared to the consensus metaschema capturing the aggregation of a simple majority of the human experts' opinions.

1. Introduction

The Semantic Network (SN) [1—3] is one of the three knowledge resources of the Unified Medical Language System (UMLS) [4,5]. SN consists of 135 broad categories called semantic types. Pairs of semantic types are connected by hierarchical (IS-A) and non-IS-A semantic relationships (in short, semantic relationships). The SN provides an overarching abstraction of the Metathesaurus (META) [6,7] which is the concept repository of the UMLS with about 900,000 concepts in [8]. The SN can help in user orientation into the large META knowledge base, as each concept in the META is categorized as belonging to one or more semantic types.

There are about 7000 semantic relationships connecting pairs of semantic types in the SN. Hence, although the size of the SN is magnitudes smaller than the size of the META, it is still hard for a user to comprehend the SN.

In order to support orientation into the SN, we introduced the notion of a *metaschema* [9]. A metaschema is a higher-level network that serves as a compact abstraction of the SN. As shown in [9], the notion of a metaschema offers various compact partial views which can help users in their orientation to the SN. In [10] an auditing technique for concept categorizations based on a metaschema was presented. These applications of metaschemas are presented in Section 2.3.

A metaschema is based on an underlying partition of the SN into connected groups of semantic types. For example, the cohesive metaschema in [9] is based on our partition of the SN presented in [11]. In our previous work, we derived the *lexical metaschema* [12] based on a lexical partition using identical word-usage among the names of semantic types and the definitions of their respective children. A more detailed description of the lexical partition and lexical metaschema is presented in Section 2.2.

In this paper we will present techniques to evaluate the lexical metaschema's quality. For this purpose, we conducted a study involving a group of experts who have published on UMLS research or related topics. In this study, each expert was asked to manually mark semantic types which are deemed as important and sufficiently different compared to their parents. These semantic types serve as roots of semantic-type groups of his/her partition of the SN. By doing this, each expert derived his/her own partition. A metaschema, called an *expert metaschema*, can then be built from each such partition. We found that these expert metaschemas vary so widely that they cannot serve as suitable evaluation yardsticks for our lexical metaschema. Therefore, we built a collection of "cumulative metaschemas," each of which represents a level of aggregation of experts' opinions. Of particular interest is the *consensus metaschema* which was selected from these cumulative metaschemas to represent a simple majority of experts' opinions. The lexical metaschema was then compared in detail to the expert metaschemas and the consensus metaschema using a statistical analysis method. The comparison results are presented and analyzed.

2. Background

2.1. Metaschema of the SN

The notion of metaschema was introduced in [9] as an abstraction of the SN. A metaschema is based on a connected partition of the SN where the SN's IS-A hierarchy is partitioned into disjoint *semantic-type groups*. A partition is said to be *connected* if each of its semantic-type groups satisfies the condition that its semantic types together with their respective IS-A links constitute a connected subgraph of the SN with a unique root. Additionally, while a semantic-type group can be a singleton (i.e., can contain only one semantic type), that singleton semantic type cannot be a leaf in the SN's hierarchy. This condition is imposed because the metaschema should manifest some size reduction, which singletons do not contribute to. However, a singleton containing a non-leaf semantic type is allowed, since it may express an important internal branching point in the metaschema.

In a metaschema, each semantic-type group of the partition is represented by a single node, called a *meta-semantic type* (MST). Two kinds of relationships connect meta-semantic types. The hierarchical *meta-child-of* relationships between