MEDINFO 2004
M. Fieschi et al. (Eds)
Amsterdam: IOS Press

Development of a Test Collection of Manually Extracted Semantic Relationships in Health Consumer Texts

Laura A. Slaughter[a] and Thomas C. Rindflesch [b]

[a]*Department of Biomedical Informatics, Columbia University, New York, New York, USA*
[b]*National Library of Medicine, Bethesda, Maryland, USA*

Abstract

Semantic relationships within knowledge bases are the links that connect concepts to one another. They are often used, for example, within information retrieval applications for search term expansion. The overall goal of this project was to manually identify the semantic relationships within health consumer question and physician-provided answer texts. We created a collection of manually identified semantic relationships for purposes of evaluating automated extraction methods. We identified a total of 509 semantic relationship instances within twelve consumer-oriented question-answer pairs (avg. of 275 words per pair). Coding of the semantic relationships was based on a set of revised relations derived from the Unified Medical Language System (UMLS) Semantic Network.

Introduction

Automated identification of semantic relationships within natural language texts is a current challenge. The semantic relationships identified were expected to be useful as a guide to further search terms. As an example, imagine a user has entered the question "Does exercise help prevent osteoporosis?" into a question-answer system and then wished to redefine the question (perhaps since an exact answer could not be determined). The relationship "exercise <*prevents*> osteoporosis" can be used as a starting point that branches out to other possible search terms and semantic relationships.

Materials and Methods

Coding of the semantic relationships was primarily based on the set of revised relations derived from the Unified Medical Language System (UMLS) Semantic Network [1]. To detect relationships, we used a thorough line-by-line microanalysis to systematically examine and interpret the question-answer texts. Coding rules were followed while manually identifying the semantic relationships to ensure consistency. For example, when selecting a semantic relation, we looked at the hierarchical structure of the relationships as well as the definitions within the Semantic Network. We always chose the most specific relationship possible. So, if the relationship was unquestionably *carries_out*, then we used that one rather than its parent *performs*. However, if we were unable to determine if it was *carries_out*, then we assigned the parent relationship *performs*. As another example of a coding rule, the values that were chosen to fill the slots could be single words or several word noun compounds. We avoided long phrases containing verbs and occasionally used nested frames.

The semantic relationships identified were recorded as frame structures. A list of slot types was defined during the process of coding. The entire test collection is represented within the frame-based Protégé-2000 Ontology Editor system[1].

Results

We identified a total of 509 semantic relationship instances. An example of a manually identified instance is illustrated in Figure 1. There were several relationships that were very frequently identified in the texts. Causal relationships made up a substantial percent of the relationships expressed in questions and in answers. Questioners, especially to explain the duration of illness, ages of individuals, number of times symptoms occur, often used time-related relationships. The relationship *treats* was repeatedly expressed in health consumer questions and also within the physician's responses. The relationship *diagnoses* appeared slightly more often in answers than in questions.

Instance of: *ingredient_of*
Text: "Chamomile tea contains an active ingredient known as apigenin," [2]
Material: apigenin
ObjectWhole: chamomile tea

Figure 1 – Example of a manually identified semantic relationship instance. The slot names are in bold text.

Conclusions

Our work has implications for important Natural Language Processing (NLP) goals: automated extraction of semantic relationships from consumer health texts. The semantic relationship instances coded from the text went through numerous iterations before arriving at a set that, although not perfect, reflects a useful representation prepared by a human. This set can be used a basis of comparison with automated attempts to extract semantic relationships from these text.

Acknowledgments

This work was supported in part by the Beta Phi Mu Doctoral Dissertation Fellowship, the Eugene Garfield Doctoral Dissertation Fellowship, and a National Library of Medicine (NLM) Training Grant.

References

[1] McCray, A., & Hole, W. (1990). The scope and structure of the first version of the UMLS Semantic Network., *Proc Annu Symp Comput Appl Med Care 1990* (pp. 126-30.).

[2] WebMD_Health (n.d.) (2001). Stress Management. my.webmd.com/content/article/3079.1680.

[1] (http://protege.stanford.edu).