Innovative Navigation of Health Discussion Forums based on Relationship Extraction and Medical Ontologies

Saeed Mohajeri, Afsaneh Esteki, Osmar R. Zaiane, and Davood Rafiei
University of Alberta, canada
zaiane@cs.ualberta.ca

Abstract

Health discussion forums are popular, contain useful information, but are difficult to navigate given the high volume of discussions and the non-standard language used. To simplify exploring the medical discussion forums, we extract from natural language text medical entities using MeSH and use three methods to find relationships between the entities, then form an ontology of those entities as an effective interface for navigating through discussions.

1. Introduction

Online health discussion forums contain a large amount of valuable user-generated contents that can be used as a useful resource to collect information. Finding a particular discussion that is relevant to a user's query is usually difficult in large discussion forums. We introduce an innovative method for navigating through discussions: A visualized version of the ontology of medical entities used in discussions helps users to see a high-level summary of the information in discussions and to find the topic they are interested in. Others have already investigated different means to extract and represent generalized terms from medical forums [1,2,3] or online articles [4] for the purpose of better organizing access to information, indicating the need of such research.

2. Methodology

We use MeSH [5], a pre-constructed ontology for the medical domain, to extract medical entities appearing in the discussions. We then use three methods for automatically extracting relationships between entities from the discussions. Finally, having a list of entities and a list of relationships between those entities, we are able to build the network and proceed with providing an efficient user interface (Figure 1) that allows navigation by presenting a big picture of related terms in the discussion and the possibility to drill through to a particular discussion or discussions.

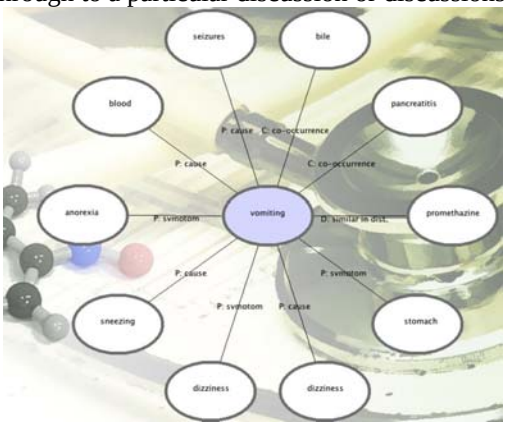

Figure 1: User Interface: Set of words related to the term "vomiting" that a user has entered.

2.1. Pattern Matching

A pattern is a sequence of words and wildcards (we refer to the wildcards as slots) that represent a certain relation. For example, *X is a type of Y* is a pattern representing the generalization relation between X and Y. We created some patterns manually by examining frequent patterns used in real medical discussions. These patterns are categorized in 4 groups: Treatments, Causes, Symptoms and Association. Table 1 shows some of our manually-written patterns. We use a pattern matching algorithm to identify relations. For each sentence we test all our patterns by mapping the tokens of the sentence to those of the pattern and calculating a matching score using an edit distance (i.e. the minimum number of edits such as insertion, deletion or replacement, needed to convert a string to another). The mapping is also based on the edit distance; words of the pattern are mapped to the most similar words in the sentence. Slots are mapped to any word but preferably medical entities listed in our

database. These matches, if highly scored, are stored and used to list relations for a given entity. For instance in Figure 1, “Vomiting” is a query used in a slot of these stored matching instances. Relations are found such as “seizure” connected by a causality relation and “dizziness” related by a symptomatic one.

Table 1: Examples of patterns.

Category	Pattern
Treatments	X is cured by Y X is used in the treatment of Y
Causes	X is a result of Y X causes Y
Symptoms	X symptoms include Y Signs of X are Y
Associations	X is a type of Y X is associated with Y

2.2. Co-Occurrence Matrix

If two entities are strongly related, they are likely to appear together in several sentences. Thus, we can estimate relations using a measure of how many times two entities co-occur in sentences of our data. To this end, we build a co-occurrence matrix to contain co-occurrence scores between pairs of entities. The score depends upon the distance in the sentence between the entities, the size of the sentence and the presence of *relational words*. The closer in the sentence two entities are, the higher their co-occurrence score. Also the shorter a sentence is the higher the score. This is because short sentences are more probable to represent coherent relations between entities than longer ones. Relational words are words frequently used for representing relations such as associated, related, causes ... The more relational words a sentence has, the higher score the co-occurrence it contains get. Depending upon the position of the relational word vis-à-vis the co-occurring entities in the sentence, the scores of X-Y and Y-X may not be incremented to the same final score and thus the matrix is not symmetric. To find entities related to a particular word, we scan the corresponding row and find the column in which their intersection gets maximized. In Figure 1, Vomiting and pancreatitis are related by co-occurrence.

2.3. Distribution-based Method

Two words are related if they tend to have a similar distribution of co-occurrences with other words. Thus, for each entity, we look for a word that has the most similar distribution of co-occurrence scores. The co-occurrence matrix is made symmetric by adding it to its transpose and the values in each row are normalized with respect to the row's maximum. The next step is to subtract a row from another. By calculating the sum of squares of the components of resulting vector, we can measure how similar or different those two rows are. It is then possible to find the most similar row for each row of the matrix, which is equivalent to finding the most similar word to another one. In Figure 1, vomiting is indicated as related to promethazine by distribution relation. Indeed it is a drug to treat or prevent vomiting.

3. Conclusion

Medical and health discussion forums are domain-specific, which means that there are coherent discussions about topics of a specific domain. This can helps in designing ontology extraction methods that are more accurate and powerful due to the domain knowledge we may have. This ontology can be used to navigate through the forum by providing a term and seeing the present related terms then drill to other related terms or through to the discussions themselves. We implemented three different ontology extraction methods and applied them on health discussion forums to extract relations between medical entities. We used about 50,000 discussions from Healthboards (healthboards.com), eHealth-Forum (ehealthforum.com) and MedHelp (medhelp.org/forums/list) and 6,000 medical terms obtained from MeSH categories [A] Anatomy, [C] Diseases, and [E] Analytical, Diagnostic and Therapeutic Techniques and Equipment. We evaluated a set of instances of the results of each method manually. All of these three methods presented acceptable performance on our dataset. In this work we only considered terms with one word. Longer medical terms exist and not considering them is a limitation. However, an extension is trivial since our pattern matching and co-occurrence matrix methods can easily be adapted to take into account composite terms.

References

[1] Sondhi, P. Shallow information extraction from medical forum data. Int. Conf. on Computational Linguistics, 2010, pp 1158–1166, Beijing, China..

[2] Chee, B., Berlin, R., and Schatz, B. Predicting adverse drug events from personal health messages. AMIA Annual Symposium Proceedings, pp 217–226, 2011.

[3] Cong, G., Wang, L., Lin, C., Song, Y., and Sun, Y. Finding question-answer pairs from online forums. In Proceedings of SIGIR, 2008.

[4] Pedro, V., Niculescu, S., and Lita, L. Okinet: Automatic extraction of a medical ontology from wikipedia. In WiKiAI08: a workshop of AAAI2008.

[5] MeSH: Medical Subject Headings, http://www.nlm.nih.gov/mesh/meshhome.html.