Building Folk UMLS: An Approach to Finding Meaning of Folk Terms in Medical Domain

Miao Chen
School of Information Studies, Syracuse University
mchen14@syr.edu

Bei Yu
School of Information Studies, Syracuse University
byu@syr.edu

Xiaozhong Liu
School of Information Studies, Syracuse University
xliu12@syr.edu

ABSTRACT
As a medical domain knowledge base, the Unified Medical Language System (UMLS) focuses on formal and professional medical terms; online health forums contain user-generated "folk terms", which can be used to complement the UMLS vocabulary. In this paper, we propose an approach to detecting folk terms from online discussions and matching their meanings to UMLS concepts. This approach makes connections between expert-built ontology and user-generated taxonomy (folksonomy) based on term distance matching using Google distance measurement. By finding meanings of user-generated folk terms, we will build what we call "folk UMLS" ontology as enrichment to the formal UMLS ontology.

Keywords
Ontology, concept extraction, ontology concept matching

1. INTRODUCTION
The UMLS ontology is a rich organized collection of medical terms and their semantic relations, covering a broad range of knowledge in the medical domain. On the one hand, the ontology contains a great number of vocabularies contributed by domain experts in a structured format; on the other hand, in the process of information use and interaction users generate folk terms, which are in unstructured format. Both the expert terms and the folk terms reflect knowledge of medical domain, while from different perspectives. Take folk phrase "later stage breast cancer" for example, it appears in user discussion board but is not listed as a concept in UMLS. In UMLS breast cancer stages is described by phrases like "stage I breast cancer" etc, which is a more rigorous way of categorizing stages. If we can combine the folk terms and the expert terms, we can build a more comprehensive knowledge base. Motivated by this goal, in this paper we introduce a new approach to building "folk UMLS", by connecting UMLS ontology and folk terms.

In this approach we need to solve two main problems: 1) identify and extract folk terms from user generated text; 2) map folk terms to the corresponding expert terms in UMLS. There have been previous trials on extracting concepts from textual data, with many of them statistically and machine learning based. Lin & Pantel (2002) presented a concept discovery method by automatically clustering words based on semantic similarities. In Maedche & Staab (2000)'s study, words in a corpus with high tf-idf values were recognized as candidate concepts. Besides statistical method, lexicon-based, rule-based, and combined methods have also been applied for concept extraction (Cohen & Hersh, 2005), Krauthammer et al. (2000) and Gaizauskas et al. (2003).

After candidate concepts are extracted from text, they need to be matched to ontologies, either to light-weighted ontologies like taxonomy or comprehensive ones like UMLS. Researchers have proposed various kinds of approaches to matching concepts to existing ontologies. Zou et al. (2003) developed an algorithm to detect UMLS concept through permuting input text and applying syntactic and semantic filters. The MetaMap project matched noun phrases to UMLS concepts based on parsing result of free text, by computing scores of distances between UMLS concepts and original text (Aronson, 2001).

Our study takes a different approach, which considers term semantic relatedness in Google database (Cilibrasi & Vitanyi, 2007), to linking ontology to folk terms using the Google distance measure. In the following sections, we will discuss how to use Google distance in our approach in details, and then address experiments and evaluation issues, and at last summarize the paper.

2. METHOD
Our approach consists of two phases. In the first phase, we take three steps to extract folk terms from online health forums:

1) Use natural language processing tool to chunk text to phrases;
2) Extract noun phrases from the phrase chunks, select the ones with high frequencies;
3) Search these noun phrases in UMLS; if a noun phrase is not an item in UMLS and is not a stop word, then it is considered as a candidate folk term.

The performance of term extraction is evaluated after the first phase. Non-expert users will be asked to judge whether the terms are medical related or not. We then proceed to the next phase to find meanings of these folk terms. The semantic distance between a folk term and a UMLS concept is defined by their Google distance. After computing its distance to all UMLS concepts, a folk term is matched to its closest UMLS concept.

The normalized semantic relatedness between two entities is computed using the Google distance in Cilibrasi and Vitanyi (2007)'s research. It is a co-occurrence based measure of term similarity. More specifically, given one webpage containing one term, Google distance measures the probability that it contains the other term (Gligorov et al., 2007). In the "later stage breast cancer" example, this is a folk noun phrase in medical forum and our goal is to project this entity to the UMLS concepts.

It costs high to compute the semantic distance between a folk term to all UMLS concepts. We take the following approach to reduce the computing load.

A folk term could be matched to a UMLS concept in two possible ways. The first one is easier by sharing the same head noun, like "flu" in both "pig flu" and "swine flu". The other one is more difficult, that is, they do not share the head noun but are semantically equivalent, e.g. "blood sugar" and "Glucose". In this study, we focus on the first matching type and leave the second matching type for future work. Below we use the "later stage breast cancer" as example to describe steps of the matching process:

1) Identify the head noun in the folk term "later stage breast cancer". We could compute the probability of each word in UMLS, like *P("cancer") > P("stage")*. Starting from the word "cancer" we find the largest possible match in UMLS. Here the largest phrase is "stage breast cancer".
2) Then find all concepts from UMLS containing the phrase as candidate matched concepts, i.e. "stage I breast cancer", "stage II breast cancer", "stage IV breast cancer", etc.
3) Compute the Google distance between the folk term and each UMLS candidate concept by using:

$$NGD(phrase, concept) = \frac{\max\{\log f(phrase), \log f(concept)\} - \log f(phrase, concept)}{\log M - \min\{\log f(phrase), \log f(concept)\}}$$

Where f(x) is the Google returned number of hits of phrase x and M is the number of total indexed pages by Google. If the target phrase and concept never occur together on the same web page, but do occur separately, the normalized Google distance (NGD) between them is infinite. If both the phrase and concept always occur together, then their NGD is zero.
4) Select the UMLS concept that has the lowest similarity with the phrase and match them. In this example folk phrase "later stage breast cancer" is matched to UMLS concept "stage IV breast cancer".

3. EXPERIMENT & EVALUATION

In our experiment, the data set comes from health forums such as WebMD, healthboards.com, and breastcancer.org, where people discuss medical related issues. We assume that people are more likely to use folk terms in such informal settings.

We will use the OpenNLP package, which is an open source natural language processing toolkit for finding noun phrases. It will detect sentence boundaries in free text and chunk sentences to phrases. The evaluation of folk term extraction will apply mechanism of precision and recall rates, as are frequently used in automatic term recognition. The precision is the rate of correctly identified terms among all identified terms, and the recall rate is the rate of correct terms in a document (Krauthammer & Nenadic, 2004). Similarly, the performance of concept mapping will also be measured by precision and recall. The first evaluation will obtain judgment from non-experts and for the second one we will ask medical experts to decide the matching performance.

4. SUMMARY

The paper proposes a new approach to discovering folk terms and connecting folk terms and professional ontologies by using Google distance measurement. By linking them, meanings of folk terms are made explicit thus can be further used in text processing tasks. It can facilitate medical document indexing by providing more related folk terms; it can be processed to Linked Open Data format to connect to other knowledge bases like UMLS; it can also provide a matching table to doctors who and patients and may enhance their communication quality. In the future, we will explore more possibilities to use the derived mapping between folk terms and formal terms as well as evaluate their usage.

5. REFERENCES

[1] Aronson, A.R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. Proceedings of AMIA Symposium 2001, 17-21.

[2] Cilibrasi, R. L., and Vitanyi, P. M. B. (2007). The Google Similarity Distance. IEEE Transactions on Knowledge and Data Engineering, 19(3), 370-383.

[3] Cohen, A.M., and Hersh, W.R. (2005). A survey of current work in biomedical text mining. Briefings in Bioinformatics, 6(1), 57-71.

[4] Gaizauskas, R., Demetriou, G., Artymiuk, P.J., & Willett, P. (2003). Protein structures and information extraction from biological texts: the PASTA system. Bioinformatics, 19(1), 135-143.

[5] Gligorov, R., Aleksovski, Z., ten Kate, W., & van Harmelen, F. (2007). Using Google distance to weight approximate ontology matches. The 17th World Wide Web Conference 2007.

[6] Krauthammer, M., and Nenadic, G. (2004). Term identification in the biomedical literature. Journal of Biomedical Informatics, 37(6), 512-526.

[7] Krauthammer, M., Rzhetsky, A., Morozov, P., & Friedman, C. (2000). Using BLAST for identifying gene and protein names in journal articles. Gene, 259(1-2), 245-252.

[8] Lin, D., and Pantel, P. Concepts discovery from text .Proceedings of the 19th International Conference on Computational Linguistics, 577-583.

[9] Maedche, A., and Staab, S. (2000). Mining ontologies from text. Proceedings of the 12t European Workshop on Knowledge Acquisition, Modeling and Management, 189-202.

[10] Unified Medical Language System (UMLS). http://www.nlm.nih.gov/research/umls/

[11] Zou, Q., Chu, W.W., Morioka, C., Leazer, G.H., & Kangarloo, H. (2003). IndexFinder: A method of extracting key concepts from clinical texts for indexing. AMIA Annual Symposium Proceedings, 763-767.