BIOINFORMATICS APPLICATIONS NOTE
Vol. 24 no. 17 2008, pages 1971–1973
doi:10.1093/bioinformatics/btn343

Databases and ontologies

Semantic reclassification of the UMLS concepts

Jung-Wei Fan* and Carol Friedman

Department of Biomedical Informatics, Columbia University, 622 W 168th St, VC5, New York, NY10032, USA

Received on April 30, 2008; revised on June 25, 2008; accepted on July 2, 2008

Advance Access publication July 13, 2008

Associate Editor: John Quackenbush

ABSTRACT

Summary: Accurate semantic classification is valuable for text mining and knowledge-based tasks that perform inference based on semantic classes. To benefit applications using the semantic classification of the Unified Medical Language System (UMLS) concepts, we automatically reclassified the concepts based on their lexical and contextual features. The new classification is useful for auditing the original UMLS semantic classification and for building biomedical text mining applications.

Availability: http://www.dbmi.columbia.edu/∼juf7002/reclassify_production

Contact: fan@dbmi.columbia.edu

Supplementary information: Supplementary data is available at http://www.dbmi.columbia.edu/∼juf7002/reclassify_production.

1 INTRODUCTION

Semantic classification is characteristic of well-organized biomedical ontologies. In cases where the ontology is associated with a comprehensive terminology, semantic classification is very helpful for text mining in that it enables the application of knowledge-based constraints during the process of extraction and interpretation. For example, by simply searching co-occurrences of textual terms belonging to a disorder class and to a gene class it is possible to extract potential disease–gene associations from text, although a more complex strategy would likely improve performance. The Unified Medical Language System (UMLS) (Lindberg *et al*., 1993) is such a representative infrastructure which combines a comprehensive terminology (the Metathesaurus) and an ontology (the Semantic Network). Synonyms are grouped into individual concepts in the Metathesaurus, and each concept is assigned one or more semantic type(s) in the Semantic Network. The semantic types have been used in tasks such as automatic annotation of enzyme classes (Hofmann and Schomburg, 2005), information extraction for pharmacogenomics (Ahlers *et al*., 2007) and discovering biological knowledge from the literature (Srinivasan and Libbus, 2004). However, questionable semantic type assignments have been observed, which compromised the performance of the applications built upon them. For example, Hofmann and Schomburg found concepts, such as 'Increased activities', were inappropriately assigned the type **Disease or Syndrome**, introducing false positives into their disease-related annotations. The UMLS fixed the problem later by assigning the concept to the broader and more neutral type **Finding**. However, we noticed that false negatives exist also. For example, many disease-related concepts such as 'Hyperkalemia' and 'Hypertriglyceridemia' are assigned **Finding** and would be missed by programs that use only **Disease or Syndrome**. The problem gets trickier as **Finding** also contains biological processes such as 'Mitotic activity' and 'Nitrogen balance', as well as many general concepts such as 'Yawning' and 'Unemployment'. Reclassification methods that can help audit those heterogeneous semantic types should benefit the associated applications.

Research has been performed in auditing the Semantic Network classification (Gu *et al*., 2004) and regrouping the semantic types into broader classes (McCray *et al*., 2001). In our previous work, we developed two automated methods to reclassify the UMLS concepts into several broad classes that are useful for biomedical text mining tasks. Both methods involve building classifiers for broad semantic classes, such as *microorganism*, *biologic function* and *gene or protein*, but each uses different features for training. One method uses contextual terms with specific syntactic information extracted from a training corpus, and computes distributional similarity for classification (Fan and Friedman, 2007). The other method implements a Naïve Bayesian (NB) classifier with bag of words prepared from the Metathesaurus strings associated with the concepts (Fan *et al*., 2007). We found the methods to be complementary to each other and suggested that both of their outputs should be considered in determining the final classification. Figure 1 illustrates the architecture, in which the example concept 'Hyperkalemia' is classified to the *disorder* class by both classifiers.

We created the reclassification database using the methods described earlier, with the following improvements: more semantic classes were added, a much larger training corpus was used for the distributional classifier and many more concepts were reclassified. A new evaluation was also performed corresponding to the latest

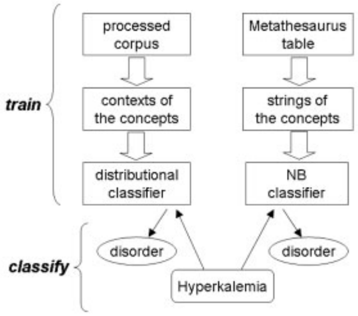

Fig. 1. Training and classifying with our two classifiers.

*To whom correspondence should be addressed.

adaptations, which is described along with details of the database implementation in the following section.

2 IMPLEMENTATION

2.1 Production

The database includes 14 broad classes: *anatomy*, *behavior*, *biologic function*, *disorder*, *gene or protein*, *geographic area*, *microorganism*, *organism*, *organization*, *population group*, *procedure*, *specialty or specialist*, *substance* and *none of above* (see Supplementary Material for composition of the broad classes). A huge syntactically and semantically processed corpus[1] (including ∼14 million PubMed abstracts) was used to generate the contextual features for the distributional classifier. The 2007AC version of the Metathesaurus MRCONSO table was used to generate the lexical features for the NB classifier. Since our reclassification aims at contributing to text mining applications, we focus on reclassifying only concepts (more specifically, those of level 0 terminologies and SNOMED-CT) that appear in the corpus. In summary, we processed 277 732 UMLS concepts with the two classifiers. Note that the distributional classifier was not applicable to 174 257 of the concepts, because no contextual features were found in the corpus. The deliverable is a flat file table with each row consisting of fields delimited by '|': UMLS concept unique identifier, the originally assigned semantic type identifier(s), the top predicted class by the NB classifier, the second predicted class by the NB classifier, the top predicted class by the distributional classifier (could be empty) and the second predicted class by the distributional classifier (could be empty). For example, C0599281 | T067 | biologic_function | gene_protein | biologic_function | substance.[2]

2.2 Evaluation

A test set of 300 concepts was randomly sampled from the 277 732 described earlier and was excluded from training. The 300 were then randomly divided into three subsets of equal size and randomly assigned to three annotators with Doctor of Medicine (MD) degrees. Each of the annotators was in charge of two subsets so that each subset was annotated by two annotators. The annotators were given the concept strings and asked to classify each concept into one or more of the broad class(es), which they considered appropriate. The third annotator was asked to perform an independent classification whenever the other two disagreed, and a scoring method based on majority votes was applied to create the gold standard (see Supplementary Material for the annotations and gold standard). The inter-annotator agreement was computed using a variant Kappa statistic that handles multiple class labels by multiple annotators (Mezzich *et al.*, 1981). The accuracies of the automatic classifiers were computed, with tied classifications counted as half correct.

3 RESULTS AND DISCUSSION

In the database, that was generated using our methods, about 50 000 of the 277 732 processed concepts were reclassified into a different broad class. The inter-annotator agreement measured by Kappa statistic was fairly high at 0.80. However, 13 of the 300 concepts did not receive a valid majority vote (i.e. there was no top class that was agreed upon by at least two of the annotators), and therefore, they were removed from the gold standard (i.e. 287 concepts were then used in evaluating our classifiers). The NB classifier achieved an accuracy of 0.78, and the distributional classifier achieved that of 0.63. By requiring the distributional classifier to be applied only to the concepts with at least five distinct contextual features, the accuracy increased to 0.77, but the coverage decreased to about 0.23. Note that the NB classifier always has 100% coverage of the concepts. The error analysis showed that it was difficult to automatically differentiate *gene or protein* from *substance*. Another pair which was difficult to differentiate was *microorganism* versus *organism*. The difficulty of resolving such broader/narrower classes was also noted by one of the annotators. A related issue is that the gold standard was not 100% correct. For example, the jumping spider 'Eris' was classified as *none of above* by two annotators, but both of our classifiers classified it as *organism*, which is supported by its original semantic type **Invertebrate**. We found several such cases indicating that our methods have the potential to outperform and complement human knowledge.

Qualitative evaluation concerning the reclassification of concepts associated with the heterogeneous semantic types showed promising results. For example, 'Inotropism' and 'Receptor internalization' were reclassified as *biologic function*, whereas the original type **Phenomenon or Process** contains diverse concepts, such as 'Traffic accidents' and 'Entropy'. The concept 'Leptospira interrogans serovar Bratislava' was reclassified as *microorganism* (confirmed to be correct based on the NCBI Taxonomy), showing that the original type **Immunologic Factor** is incorrect. The **Organic Chemical** 'Lys-Lys' was reclassified as *gene or protein*, which is supported by its parent concept 'Dipeptides', assigned to **Amino Acid, Peptide or Protein**; this example shows that our reclassification helps to cross-validate another rule-based auditing method, which requires that a child concept should always be assigned a semantic type that is not broader than that of its parent concepts. In conclusion, the results demonstrated that our methods can assist manual auditing or be integrated with other automatic methods, and that the reclassifications should improve the recall and precision of applications (e.g. information extraction or semantics-based parsing) that use the semantic classification of the UMLS concepts.

ACKNOWLEDGEMENTS

We thank Drs Peter Hung, Herbert Chase and Daniel Stein for helping create the gold standard.

Funding: This work was supported by Grants R01 LM7659 and R01 LM8635 from the National Library of Medicine of USA.

Conflict of Interest: none declared.

REFERENCES

Ahlers,C.B. *et al.* (2007) Extracting semantic predications from MEDLINE citations for pharmacogenomics. *Pac. Symp. Biocomput.*, **12**, 209–220.

Fan,J.W. and Friedman,C. (2007) Semantic classification of biomedical concepts using distributional similarity. *J. Am. Med. Inform. Assoc.*, **14**, 467–477.

Fan,J.W. *et al.* (2007) Using contextual and lexical features to restructure and validate the classification of biomedical concepts. *BMC Bioinformatics*, **8**, 264.

[1] The 2005 database of the MEDLINE/PubMed Baseline Repository (MBR) http://mbr.nlm.nih.gov/.

[2] The strings of the concepts are not displayed to avoid potential copyright issues, but licensed users can always trace them through the identifiers.

Gu,H.H. *et al*. (2004) Auditing concept categorizations in the UMLS. *Artif. Intell. Med*., **31**, 29–44.

Hofmann,O and Schomburg,D. (2005) Concept-based annotation of enzyme classes. *Bioinformatics*, **21**, 2059–2066.

Lindberg,D.A. *et al*. (1993) The Unified Medical Language System. *Methods Inf. Med*., **32**, 281–291.

McCray,A.T. *et al.* (2001) Aggregating UMLS semantic types for reducing conceptual complexity. *Medinfo*, **10**, 216–220.

Mezzich,J.E. *et al*. (1981) Assessment of agreement among several raters formulating multiple diagnoses. *J. Psychiatr. Res*., **16**, 29–39.

Srinivasan,P and Libbus,B. (2004) Mining MEDLINE for implicit links between dietary substances and diseases. *Bioinformatics*, **20**(Suppl. 1), i290–i296.