Automatic inference of indexing rules for MEDLINE

Aurélie Névéol and Sonya E. Shooshan
National Library of Medicine
8600 Rockville Pike
Bethesda, MD 20894, USA
{neveola,sonya}@nlm.nih.gov

Vincent Claveau
IRISA - CNRS
Campus de Beaulieu
35042 Rennes, France
Vincent.Claveau@irisa.fr

Abstract

This paper describes the use and customization of Inductive Logic Programming (ILP) to infer indexing rules from MEDLINE citations. Preliminary results suggest this method may enhance the subheading attachment module of the Medical Text Indexer, a system for assisting MEDLINE indexers.

1 Introduction

Indexing is a crucial step in any information retrieval system. In MEDLINE®, a widely used database of the biomedical literature, the indexing process involves the selection of Medical Subject Headings (MeSH®) in order to describe the subject matter of articles. The need for automatic tools to assist human indexers in this task is growing with the increasing number of publications in MEDLINE. The Medical Text Indexer (MTI) (Aronson et al., 2004) has been available at the U.S. National Library of Medicine (NLM) since 2002 to provide indexers with MeSH main heading recommendations (e.g. *Aphasia*, *Patient Care*...) when they create MEDLINE citations. This paper describes a method to enhance MTI with the capacity to attach appropriate MeSH subheadings (e.g. *metabolism*, *pharmacology*) to these main headings in order to provide MeSH pair recommendations (e.g. *aphasia/metabolism*), which are more specific and therefore a significant asset to NLM indexers.

Subheading attachment can be accomplished using indexing rules such as:

```
If a main heading from the "Anatomy"
tree and a "Carboxylic Acids" term are
recommended for indexing, then the pair
"[Carboxylic Acids]/pharmacology" should
also be recommended.
```

Sets of manual rules developed for a few subheadings show good precision but low recall. The development of new rules is a complex, time-consuming task. We investigate a novel approach adapting Inductive Logic Programming (ILP) to the context of MEDLINE, which requires efficient processing of large amounts of data.

2 Use of Inductive Logic Programming

ILP is a supervised machine learning technique used to infer rules that are expressed with logical clauses (Prolog clauses) based on a set of examples also represented using Prolog. A comprehensive description of ILP can be found in (Muggleton and Raedt, 1994). We selected this method because it is able to provide simple representations for relational problems and produces rules that can be easily interpreted. One caveat to the use of ILP is the complexity of rule inference from large sets of positive and negative examples. Considering each of the 24,000 MeSH main headings independently would not be computationally feasible. For this reason, based on work by Buntine (1988) we introduce a new definition of subsumption that allows us to go through the set of examples efficiently by exploiting hierarchical relationships between main headings. This type of subsumption is in fact suitable for any rule inference problem involving structured knowledge encoded by ontologies.

BioNLP 2008: Current Trends in Biomedical Natural Language Processing, pages 88–89,
Columbus, Ohio, USA, June 2008.

Subheading	Method	Nb. rules	Precision (%)	Recall(%)	F-measure(%)
Overall	ILP	587	47	**32**	**38**
	Manual	69	**59**	10	18
	Baseline	-	32	11	16

Table 1: Performance on the test corpus using MTI main heading recommendations

3 Experiments

ILP rules were induced using a training corpus of 100,000 citations randomly chosen from MEDLINE 2006. Another corpus of 100,000 MEDLINE 2006 citations was used for testing. ILP rules were applied on the test corpus using main headings automatically retrieved by MTI as triggers. The performance of ILP was compared to manual rules and a baseline consisting of randomly formed pairs according to their distribution in MEDLINE prior to 2006. Overall results obtained on 4 subheadings are presented in Table 1.

4 Discussion

Performance. As expected, the use of MTI to produce main heading recommendations used as triggers for the rules results in comparable precision but lower recall compared to the theoretical assessment. In spite of this, the performance obtained by ILP rules is superior to the baseline and shows the best F-measure. The precision obtained by the manual rules, when they exist, is higher, but they produce a recall inferior to ILP and even to the baseline method.

ILP *vs.* manual rules. A detailed analysis of the rules obtained shows that not all ILP rules are easily understood by indexers. This is due to some unexpected regularities which do not seem to be relevant but nonetheless achieved good results on the training data used to infer rules.

Furthermore, we noticed that while most rules typically contain a "trigger term" (e.g. *Anatomy* in our previous example) and a "target term" (e.g. *Carboxylic Acids* above), in some ILP rules the target term can also serve as the trigger term. Some changes in the ILP inferring process are foreseen in order to prevent the production of such rules.

Rule filtering *vs.* manual review. Preliminary experiments with producing ILP rules suggested that improvement could be achieved by 1/ filtering out rules that showed a comparatively low precision on the training corpus when applied to main headings retrieved by MTI; and 2/ by having an indexing expert review the rules to improve their readability. On most subheadings, filtering had little impact but generally tended to improve precision while F-measure stayed the same, which was our goal. The manual review of the rules seemed to degrade the performance obtained with the original ILP.

5 Conclusion and perspectives

We have shown that ILP is an adequate method for automatically inferring indexing rules for MEDLINE. Further work will be necessary in order to obtain rules for all 83 MeSH subheadings. Subsequently, the combination of ILP rules with other subheading attachment methods will be assessed. We anticipate that the rule sets we have obtained will be integrated into MTI's subheading attachment module.

Acknowledgments

This study was supported in part by the Intramural Research Programs of NIH, NLM. A. Névéol was supported by an appointment to the NLM Research Participation Program administered by ORISE through an inter-agency agreement between the U.S. Department of Energy and NLM.

References

Alan R. Aronson, James G. Mork, Clifford W. Gay, Susanne M. Humphrey, and Willie J. Rogers. 2004. The NLM Indexing Initiative's Medical Text Indexer. In *Proceedings of Medinfo 2004*, San Francisco, California, USA.

Wray L. Buntine. 1988. Generalized Subsumption and its Application to Induction and Redundancy. *Artificial Intelligence*, 36:375–399.

Stephen Muggleton and Luc De Raedt. 1994. Inductive logic programming: Theory and methods. *Journal of Logic Programming*, 19/20:629–679.