MetaMap is a superior baseline to a standard document retrieval engine for the task of finding patient cohorts in clinical free text

K. Bretonnel Cohen
Computational Bioscience Program
U. Colorado School of Medicine
and
Department of Linguistics
U. of Colorado at Boulder
kevin.cohen@gmail.com

Tom Christiansen
Comput. Bioscience Prog.
U. Colorado Sch. of Medicine
tchrist@perl.com

Lawrence E. Hunter
Computational Bioscience Program
U. Colorado School of Medicine
larry.hunter@ucdenver.edu

Abstract

The goal of this work was to establish a reasonable baseline for research in patient cohort retrieval from clinical free text. Much recent work has used Lucene for this purpose. Our approach was to use MetaMap alone. We found that although many TREC 2011 Electronic Medical Records track participants found it difficult to beat a Lucene baseline, our MetaMap-based baseline did outperform a number of Lucene runs. We propose that MetaMap is a more valid baseline than Lucene, providing essential concept extraction, and that failure to make use of this industry-standard tool results in an unfairly low baseline for evaluation of system outputs.

1 Introduction

The TREC 2011 Electronic Medical Records track involved the task of retrieving records of patients belonging to a particular cohort for comparative effectiveness research, given a set of natural language queries and a large collection of free-text clinical documents (Voorhees and Tong, 2011). Many teams used Lucene as a baseline for evaluating their systems, and found that even out-of-the-box, it could be difficult to beat.

MetaMap (Aronson, 2001) is the primary tool for detecting mentions of clinical concepts in text, and is very widely used for that purpose. We felt that in the light of the task appropriateness and free availability of this resource, it is the most reasonable baseline for efforts like the TREC 2011 Electronic Medical Records task. Therefore, we built a patient record retrieval system based entirely on this tool.

2 Methods

We used MetaMap in its default configuration to index the collection of clinical records. We used Perl to create simple hashs that mapped the CUIs that are output by MetaMap ("Concept Unique Identifiers," each belonging to a specific biomedical concept in the UMLS Metathesaurus) to patient records. Only CUIs with 1.000 certainty were included. We repeated the process with MetaMap's NegEx (Chapman et al., 2001a; Chapman et al., 2001b) option enabled. This simple approach resulted in two hashes—one for non-negated concepts in clinical records, and one for negated concepts in patient records.

At run time, we ran MetaMap on the queries, with the NegEx option enabled. For each set of CUIs output by MetaMap from the queries, we used the hashes to retrieve any patient records that contained mentions of the full set of CUIs from that query.

Like the majority of runs, our single run was unscored.

3 Results

Although many teams found it difficult to outperform Lucene-based baselines as measured with the bpref metric, the MetaMap-based approach did outperform a number of Lucene-based runs.

4 Discussion

We conclude that MetaMap provides a more stringent baseline for patient cohort retrieval from free text than does Lucene. Although Lucene is an obvious and easily implementable baseline approach,

MetaMap is freely available, easy to implement, and the most widely accepted tool for the task of finding clinical concepts, and therefore it is not valid to ignore it as a baseline. Its performance in a very simple application like ours does not produce high results (although the highest-performing team did use MetaMap as an integral part of their system (Demner-Fushman et al., 2011)), but its output provides results sufficiently better than Lucene to make it a more stringent, and therefore more indicative and scientifically appropriate, baseline.

Acknowledgments

We gratefully acknowledge the assistance of Lan Aronson and his team at the National Library of Medicine for assistance with running MetaMap. This research was funded by grants 2 R01 LM009254-06 and 5 R01 LM008111-07 from the National Institutes of Health to Lawrence E. Hunter.

References

A. Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. In *Proc AMIA 2001*, pages 17–21.

W.W. Chapman, W. Bridewell, P. Hanbury, G.F. Cooper, and B. Buchanan. 2001a. Evaluation of negation phrases in narrative clinical reports. In *Proceedings of the AMIA Symposium 2001*, pages 105–109.

W.W. Chapman, W. Bridewell, P. Hanbury, G.F. Cooper, and B.G. Buchanan. 2001b. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 34:301–310.

D. Demner-Fushman, S. Abhyankar, A. Jimeon-Yepes, R. Loane, B. Rance, F. Lang, N. Ide, E. Apostolova, and A.R. Aronson. 2011. A knowledge-based approach to medical records retrieval. In *Text Retrieval Conference (TREC 2011) Proceedings*, pages 163–172.

Ellen M. Voorhees and Richard M. Tong. 2011. Overview of the TREC 2011 medical records track (draft). In *Text Retrieval Conference (TREC 2011) Proceedings*, pages 94–98.