The Evolution of MetaMap, a Concept Search Program for Biomedical Text

Alan R. Aronson, PhD and François M. Lang, MSE
alan@nlm.nih.gov, flang@mail.nih.gov

Lister Hill National Center for Biomedical Communications (LHNCBC)
U.S. National Library of Medicine, Bethesda, MD 20894

Abstract

MetaMap[1] is a widely available program providing access to the concepts in the UMLS® Metathesaurus® from biomedical text. MetaMap arose in the context of an effort to improve biomedical text retrieval, specifically retrieval of MEDLINE®/PubMed® citations.[2,3] It provided a link between the text of biomedical literature and the knowledge, including synonymy relations, embedded in the Metathesaurus. Early MetaMap development was guided by linguistic principles which provided both a rigorous foundation and a flexible architecture. After some experience with MetaMap, it became clear that it could be applied to tasks other than retrieval, including text mining,[4] classification, question answering,[5] knowledge discovery,[6] and concept-based indexing.[7]

USES OF METAMAP

MetaMap has been used by NLM researchers and outside users since 1994 and is currently available via web access, a downloadable Java implementation (MMTx), an Application Programming Interface (API), and most recently, a downloadable version of the complete Prolog implementation of MetaMap itself. At NLM, MetaMap has analyzed the equivalent of the entire MEDLINE corpus (titles/abstracts) many times over. Batch MetaMap[8] can analyze several thousand typical citations per hour.

RESEARCH-DRIVEN DEVELOPMENT

Virtually all MetaMap development has been driven by issues arising naturally from research efforts. Some of these issues are theoretically substantial and deep; others are practical and straightforward. They include tokenization issues, output formats, issues relating to text genre or application task, and algorithm tuning.

CONCLUSION

MetaMap has evolved significantly with regard to function, implementation and distribution vehicles since its inception in the mid 1990s, and it is currently used by many groups in the biomedical informatics community. Some of the near-term plans for further MetaMap development include

- adding chemical name recognition to MetaMap's higher-order tokenization capabilities; and
- enhancing MetaMap's WSD accuracy by adding more WSD algorithms and basing final ambiguity resolution on a voting mechanism.

Acknowledgements

This work was supported by the Intramural Research Program of the NIH, National Library of Medicine.

References

1. Aronson AR. Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. *Proc AMIA Symp.* 2001:17-21.

2. Aronson AR, Rindflesch TC, Browne AC. Exploiting a Large Thesaurus for Information Retrieval. *Proc RIAO 94*:197-216, 1994.

3. Aronson AR, Rindflesch TC. Query Expansion using the UMLS Metathesaurus. *Proc AMIA Annu Fall Symp.* 1997:485-9.

4. Ahlers CB, Fiszman M, Demner-Fushman D, Lang FM, Rindflesch TC. Extracting semantic predications from Medline citations for pharmacogenomics. *Pac Symp Biocomput.* 2007:209-20.

5. Demner-Fushman D, Humphrey SM, Ide NC, Loane RF, Mork JG, Ruch P, Ruiz ME, Smith LH, Wilbur WJ, Aronson AR. Combining resources to find answers to biomedical questions. *Proc TREC* 2007, 205-14.

6. Weeber M, Klein H, Aronson AR, Mork JG, Jong-Van Den Berg L, Vos R. Text-Based Discovery in Biomedicine: The Architecture of the DAD-system. *Proc AMIA Symp.* 2000:903-7.

7. Aronson AR, Mork JG, Gay CW, Humphrey SM, Rogers WJ. The NLM Indexing Initiative's Medical Text Indexer. *Stud Health Technol Inform.* 2004;107(Pt 1):268-72.

8. http://skr.nlm.nih.gov/batch-mode