MetaMap in the CALBC Workshop II

James Mork Lee Peters Antonio Jimeno-Yepes
Alan R. Aronson Olivier Bodenreider

National Library of Medicine
8600 Rockville Pike, Bethesda, 20894, MD, USA

1 Introduction

MetaMap [4] is a tool which maps biomedical text to UMLS® Metathesaurus® concepts. In MetaMap, input text undergoes a lexical/syntactic analysis consisting of a first analysis in which tokens, sentence boundaries and acronyms or abbreviations are identified and each token is assigned a part of speech. Input words are mapped to the SPECIALIST lexicon [8] using lexical lookup and then the SPECIALIST minimal commitment parser [8] identifies phrases and their lexical heads. The identified phrases are processed to generate variants (normally by table lookup), then candidates (Metathesaurus strings) are identified computing and evaluating their match to the input text. Then mapping constructions are produced in which candidates found in the previous step are combined and evaluated to produce a final result that best matches the phrase text. Finally word sense disambiguation (WSD) might optionally be used, in which mappings involving concepts that are semantically consistent with surrounding text are favored. MetaMap is available from [2]. Downloads are restricted and require a valid UMLSKS user account. A 2011 MEDLINE baseline annotation with MetaMap is available from [3].

2 Methods

From the two available CALBC sets, we have performed the annotation of the 175K citations set. MetaMap has been configured to run using the 2010AA version of the UMLS. Preprocessing of the UMLS Metathesaurus is described in [6].

Several customizations of MetaMap were required to adjust to the requirements from the challenge guideline [1]. MetaMap has been adapted to run on the sentence boundary annotation provided in the citation set. The output of MetaMap has been turned into the IeXML format. The 2010 version of the UMLS Semantic Network has been mapped back to an earlier version specified in the CALBC guidelines, in which recent changes to the Organism hierarchy of the Semantic Network are not reflected. In addition, CALBC also uses its own mappings between Semantic Types and Semantic Groups. A mapping table has been generated to produce the set of Semantic Group annotations required in the guidelines.

MetaMap is a highly configurable tool. We have produced three runs of increasing complexity with the following options:

1. MetaMap with default options (metamap10 -Z 10 -% format -E). The default options use the strict model of the Metathesaurus [6]. This model should produce the highest precision of the annotation when compared to the moderate or relaxed model. It contains 2,424,017 (44.99%) of the 5,394,495 English Metathesaurus strings.

2. MetaMap with default options combined with Word Sense Disambiguation (WSD) (metamap10 -Z 10 -y -% format -E). Ambiguity is a main concern since in the UMLS Metathesaurus [7]. The WSD algorithm used in MetaMap is based on the JDI method [5].

3. MetaMap with default options combined with WSD and quick composite phrases (metamap10 -Z 10 -yQ -% format -E). A composite phrase is a simple phrase followed by any prepositional phrase optionally followed by one or more of prepositional phrases. An example is "pain on the left side of the chest" which will map to "Left sided chest pain" rather than separate concepts as it would without the option. The quick composite phrases option is still experimental.

References

[1] CALBC challenge II guidelines. http://www.ebi.ac.uk/Rebholz-srv/CALBC/challenge_guideline.pdf.

[2] MetaMap website. http://metamap.nlm.nih.gov.

[3] NLM LHNCBC Semantic Knowledge Representation website. http://skr.nlm.nih.gov.

[4] A.R. Aronson and F.M. Lang. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229, 2010.

[5] S.M. Humphrey, W.J. Rogers, H. Kilicoglu, D. Demner-Fushman, and T.C. Rindflesch. Word sense disambiguation by selecting the best semantic type based on Journal Descriptor Indexing: Preliminary experiment. *Journal of the American Society for Information Science and Technology*, 57(1):96, 2006.

[6] F.M. Lang and A.R. Aronson. Filtering the UMLS Metathesaurus for MetaMap. http://skr.nlm.nih.gov/papers/references/filtering10.pdf.

[7] F.M. Lang, S.E. Shooshan, J.G. Mork, and A. R. Aronson. Ambiguity in the UMLS Metathesaurus. http://skr.nlm.nih.gov/papers/references/ambiguity10.pdf.

[8] A.T. McCray, A.R. Aronson, A.C. Browne, T.C. Rindflesch, A. Razi, and S. Srinivasan. UMLS knowledge for biomedical language processing. *Bulletin of the Medical Library Association*, 81(2):184, 1993.