DEFINDER: Rule-Based Methods for the Extraction of Medical Terminology and their Associated Definitions from On-line Text

Judith L. Klavans Ph.D.[1], Smaranda Muresan[2]
[1]Center for Research on Information Access, Columbia University
[2]Department of Computer Science, Columbia University

INTRODUCTION

The problem addressed in this paper concerns the automatic identification and extraction of medical terms along with their definitions and modifiers from full text consumer-oriented medical articles. The system, `DEFINDER` (Definition Finder), uses rule-based techniques. The output of our system can be used in several applications: creation and/or enhancement of on-line terminological resources, summarization and text categorization according to level of expertise, e.g. lay vs. technical.

METHOD

MEDLINEplus, the MEDLINE equivalent for consumer health information, was used as the seed corpus, and http://www.cardio.com/articles.html was selected since it was well-edited and structured, and thus suited to rule-based pattern extraction techniques. From this, a corpus was built and randomly split using 75% for development and 25% for testing. Nearly 60% of the definitions are introduced by a limited set of text markers ('--', '()'), the other 40% being identified by more complex linguistic phenomena (anaphora, apposition, conjoined definitions). Based on this analysis, a two-stage system was developed containing: (1) a pattern extractor using text markers (consisting of a tagger and a finite state grammar); and (2) a natural language parser, English Slot Grammar[1] for more complex linguistic structures. We used the UMLS as domain knowledge.

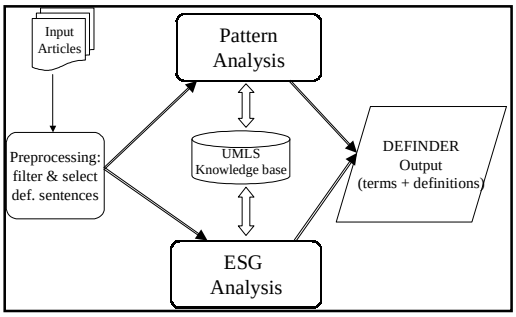

Figure1 DEFINDER architecture

Examples of output are:

(1) *foam cells* - white blood cells that have ingested fat. (Found in both UMLS and OMD)
(2) *homocysteine* - a byproduct of the metabolism of methionine, an essential aminoacid found in meat and dairy products. (In UMLS, not OMD)

RESULTS

Results were evaluated using two methods, to reflect the quality of (1) terms alone, or (2) terms and definitions. A base test set of 93 terms and their associated definitions was identified. For terms alone, 84% precision and 83% recall was achieved with DEFINDER. The second evaluation considered DEFINDER results against three existing on-line dictionaries and glossaries: UMLS, On-line Medical Dictionary (http://www.graylab.ac.uk/omd/) and Glossary of Popular and Technical Medical Terms (http://allserv.rug.ac.be/%7Ervdstich/eugloss/welcome.html).

	UMLS	OMD	Glossary
defined	60%(56)	76%(71)	21.5%(20)
undefined	24%(22)	-	-
absent	16%(15)	24%(22)	78.5%(73)

Table 1 Comparison of term extraction with three existing on-line dictionaries

Table 1 shows that on-line medical dictionaries appear to be incomplete, although an in-depth analysis of absent and undefined terms [2] indicates that partial matches must be considered.

FUTURE DIRECTIONS

Our initial study has raised a number of questions which we are currently addressing: (1) extending the rule set to embrace additional patterns thus improving recall while maintaining high precision; (2) analyzing additional corpora; and (3) further examining merging techniques.

[1] McCord M.C . The Slot Grammar System. Research Report; IBM Research Division, T.J. Watson Research Center; 1991.
[2] McCray A.T, Browne A.C. Discovering the Modifiers in a Termonology Data Set. AMIA Annual Symposium. 1998 on CD-ROM [D004985]