Evaluation of DEFINDER: A System to Mine Definitions from Consumer-oriented Medical Text

Judith L. Klavans
Center for Research on Information Access
Columbia University
New York, NY, 10027
klavans@cs.columbia.edu

Smaranda Muresan
Department of Computer Science
Columbia University
New York, NY, 10027
smara@cs.columbia.edu

ABSTRACT

In this paper we present DEFINDER, a rule-based system that mines consumer-oriented full text articles in order to extract definitions and the terms they define. This research is part of Digital Library Project at Columbia University, entitled PERSIVAL (PErsonalized Retrieval and Summarization of Image, Video and Language resources) [5]. One goal of the project is to present information to patients in language they can understand. A key component of this stage is to provide accurate and readable lay definitions for technical terms, which may be present in articles of intermediate complexity.

The focus of this short paper is on quantitative and qualitative evaluation of the DEFINDER system [3]. Our basis for comparison was definitions from Unified Medical Language System (UMLS), On-line Medical Dictionary (OMD) and Glossary of Popular and Technical Medical Terms (GPTMT). Quantitative evaluations show that DEFINDER obtained 87% precision and 75% recall and reveal the incompleteness of existing resources and the ability of DEFINDER to address gaps. Qualitative evaluation shows that the definitions extracted by our system are ranked higher in terms of user-based criteria of usability and readability than definitions from on-line specialized dictionaries. Thus the output of DEFINDER can be used to enhance existing specialized dictionaries, and also as a key feature in summarizing technical articles for non-specialist users.

Keywords

Text data mining, medical digital libraries, natural language processing, automatic dictionary creation.

1. The Digital Library and Text Mining for Definitions: the DEFINDER System

The existence of massive digital libraries containing freeform documents has created an unprecedented opportunity to develop and apply effective and scalable text mining techniques for the automatic extraction of knowledge from unstructured text [1]. Text mining applications raise particularly challenging problems within digital libraries since they involve large collections of unstructured documents. Our approach is to combine shallow natural language processing techniques with deep grammatical analysis in order to efficiently mine text.

Automatic identification and extraction of terms from text has been widely studied in the computational linguistics literature [2], and many systems exist for this task using both symbolic and statistical techniques. The extraction of definitions and their associated terms has been less widely studied, although extraction of lexical knowledge has a rich literature [7].

Through an analysis of a set of consumer-oriented medical articles, we identified typical cue-phrases and structural indicators that introduce definitions and the defined terms. Our system, DEFINDER, is based on two main functional modules: 1) a shallow text processing module which performs pattern analyses using a finite state grammar, guided by cue-phrases ("is called", "is the term used to describe", "is defined as", etc.) and a limited set of text-markers ((), --) and 2) a grammar analysis module that uses a rich, dependency-oriented lexicalist grammar (English Slot Grammar [4]) for analyzing more complex linguistic phenomena (e.g. apposition, anaphora).

JCDL '01, June 24-28, 2001, Roanoke, Virginia.

2. Evaluation: Users and Uses

In this brief paper, we present the results of three methods to evaluate the output of our system: 1) performance in terms of precision and recall, 2) quality of extracted definitions in terms of user-based criteria of readability, usefulness and completeness and 3) a method to evaluate the coverage of on-line specialized dictionaries. For the first two, we performed a user-centered evaluation using non-specialist subjects. For the latter we chose a set of defined terms extracted by our system and compared them against three on-line dictionaries. The results we have obtained were run over a limited set of articles in order to thoroughly test our methods before moving to a larger scale user-based evaluation of significantly more data. We present the results of three experiments to quantitatively and qualitatively measure DEFINDER output.

2.1 Definition Extraction Performance

The purpose of this experiment was to measure the performance of DEFINDER in terms of precision and recall against a human-determined "gold standard". Four subjects unrelated to the project were provided with a set of nine patient-oriented articles

and were asked to annotate definitions and the terms they define. We chose several genres (medical articles, newspapers, manual chapters, book chapters) from trusted resources. The resulting gold standard was determined by those definitions marked-up by at least 3 out of the 4 subjects and consisted of 53 definitions. DEFINDER identified 40 out of these 53 definitions obtaining 86.95% precision and 75.47% recall.

2.2 User Judgements on Definition Quality

In this experiment we asked users to rank definitions to determine if they are readable, useful or complete. The motivation is that there is unlikely to exist a single definition suitable for both specialists and non-specialists. Indeed, specialized on-line dictionaries, while valuable resources, can be too technical for non-specialists. We evaluated the quality of DEFINDER output in comparison with two specialized on-line dictionaries (UMLS and OMD). Eight subjects not qualified in the medical domain participated in the experiment. They were provided with a list of 15 randomly chosen medical terms and their definitions from these three sources. The task was to assign to each definition a quality rating for three criteria: usefulness (U), readability (R) and completeness (C) on a scale of 1 to 7 (1 worst, 7 best). The source of each definition was not given in order not to bias the experiment. Statistical significance tests were performed for subjects and terms using Kendall's coefficient of concordance, W [6] and the sign test [6].

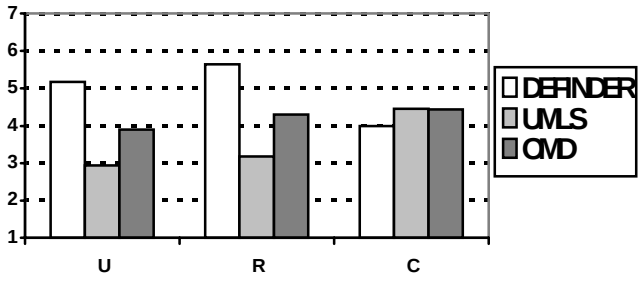

Figure 1 - Average quality rating (AQR)

We first measured the average quality rating for each of the three sources on the three criteria. The results in Fig. 1 show that DEFINDER clearly outperforms the specialized dictionaries for usefulness and readability to a statistically significant degree, given by the sign test (p=0.0003). In terms of completeness both UMLS and OMD performed slightly better (p=0.04).

One question that arises in computing the AQR is whether the high scores given by one subject can compensate for the lower values given by other subjects, thus introducing noise. To validate our results, we performed a second analysis to evaluate the relative ranking of the three definitional sources. Using Kendall's coefficient of correlation, W, we first measured the interjudge reliability on each term, and for terms with significant agreement we compute the level of correlation between them. If W was significant, we compared the overall mean ranks of the three sources. We obtained statistically significant W values for usefulness and readability (W=0.54 and W=0.45 at p=0.01 and p=0.05 respectively), while for completeness the correlation was not statistically significant. Thus Figure 2 shows the results for usefulness (U) and readability (R) for which DEFINDER outranked both UMLS and OMD.

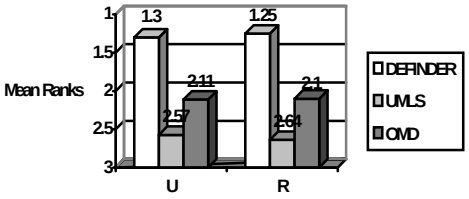

Figure 2 – Ranking

2.3 Coverage of On-line Dictionaries

DEFINDER identifies terms and definitions lacking from existing resources. To evaluate coverage, we choose a base test set of 93 terms and their associated definitions, extracted by our system from text. Three cases were found, as shown in Table 1: (1) the term is listed in one of the on-line dictionaries and is defined in that dictionary (defined); (2) the term is listed in one of the on-line dictionaries but does not have an associated definition (undefined); (3) the term is not listed in one of the on-line dictionaries (absent).

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

Table 1 Coverage of Existing Online Dictionaries

Table 1 shows that on-line medical dictionaries are incomplete compared to potential DEFINDER output. For example, column two shows that in OMD only 71 terms out of 93 are listed, thus leading to 76% completeness, while GPTMT, a glossary addressed to non-specialists is far from being complete, i.e. only 20 out of 93 terms were present. This proves the ability of DEFINDER to enhance on-line dictionaries with readable and useful definitions.

3. References

[1] Hearst, M. Untangling Text Data Mining. *Proc. of ACL'99* University of Maryland, June 20-26, 1999 (invited paper).

[2] Justeson, J. and Katz, S. Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text. Natural Language Engineering. Vol 1(1). 1995. pp. 9-27.

[3] Klavans J.L., Muresan S. DEFINDER: Rule-Based Methods for the Extraction of Medical Terminology and their Associated Definitions from On-line Text. *Proc of AMIA 2000*; pp. 1906.

[4] McCord M.C. The Slot Grammar system. IBM Report; 1991.

[5] McKeown K.R et al. PERSIVAL, a System for Personalized Search and Summarization over Multimedia Healthcare Information. *Proc of JCDL 2001*.

[6] Siegal, S. and Castellan, N.J. (1988). Non-parametric statistics for the behavioural sciences (2nd Edition). New York: McGraw Hill.

[7] Zweigenbaum P, Bouaud J, Bachimont B, Charlet J, Seroussi B, Boisvieux JF. From Text to Knowledge: a Unifying Document-Oriented View of Analyzed Medical Language. Proceedings of IMIA WG6. 1997.