Evaluation of the DEFINDER System for Fully Automatic Glossary Construction

Judith L. Klavans PhD[1], Smaranda Muresan[2]
[1]Center for Research on Information Access, Columbia University
[2]Department of Computer Science, Columbia University
New York, New York 10027

ABSTRACT

In this paper we present a quantitative and qualitative evaluation of DEFINDER, a rule-based system that mines consumer-oriented full text articles in order to extract definitions and the terms they define. The quantitative evaluation shows that in terms of precision and recall as measured against human performance, DEFINDER obtained 87% and 75% respectively, thereby revealing the incompleteness of existing resources and the ability of DEFINDER to address these gaps. Our basis for comparison is definitions from on-line dictionaries, including the UMLS Metathesaurus. Qualitative evaluation shows that the definitions extracted by our system are ranked higher in terms of user-centered criteria of usability and readability than are definitions from on-line specialized dictionaries. The output of DEFINDER can be used to enhance these dictionaries. DEFINDER output is being incorporated in a system to clarify technical terms for non-specialist users in understandable non-technical language.

INTRODUCTION

The problem we address is that on-line dictionaries of technical terms are difficult to build and are often lacking in completeness. For example in the UMLS (Unified Medical Language System) Metathesaurus,[1] basic terms are often undefined since professionals are assumed to know them. Common terms and their definitions are also missing from on-line glossaries, such as the OMD (Online Medical Dictionary) .[2] The contribution of this research is in utilizing natural language processing techniques to mine text for embedded definitions and to extract these definitions along with their associated terms. These results are used to create a new glossary for lay users and to enhance existing resources.

Our definition finding system, DEFINDER, is being developed as part of Columbia University's medical digital library project PERSIVAL [1]. At the larger level, this research is part of an effort to build useful lexical resources from existing on-line material using robust language technologies.

This paper reports on three research contributions:

1. The first quantitative evaluation of our results shows that in terms of precision and recall, measured against human performance, our system achieves 86.95% precision and 75.47% recall. These results correlate with the fact that consumer-oriented documents from trusted resources are a rich source of definitions, and demonstrate that our system offers a novel and useful technology to identify and extract terminology and definitions.
2. The qualitative evaluation of results in terms of the user-based criteria of readability, usefulness and completeness indicates that the definitions extracted from text are often more readable and useful than those in existing thesauri and glossaries.
3. A second quantitative evaluation shows that DEFINDER extracted definitions not found in existing on-line dictionaries. Thus the output from running DEFINDER over full text can be used to fill gaps in these dictionaries.

BACKGROUND

One goal of the PERSIVAL project is to present information to patients in language they can understand. A key component is to provide accurate and readable lay definitions for technical terms, which may be present in articles of intermediate complexity [2]. A preliminary version of DEFINDER, a system for automatically identifying and extracting definitions from consumer-oriented text, was presented in [3].

In order to extract valuable definitions for the non-specialist user we need reliable resources. We

[1] http://umlsks.nlm.nih.gov/
[2] http://www.graylab.ac.uk/omd/index.html

started with MEDLINEplus, the MEDLINE equivalent for consumer health information. We identified five full-text resources from different text genres that are addressed to lay people[3], as judged by a medical specialist. Definitions extracted by DEFINDER include:

TNF-alpha - *one of a class of proteins called cytokines, which allow cells to communicate with each other*
AV node - *the structure that governs impulse traffic from the atria to the ventricles*
mitral valve prolapse - *a condition in which the valve between the upper and lower chambers on the left side of the heart closes imperfectly.*

The first two definitions were judged accurate by two medical consultants, although the third was judged possibly vague. (We discuss accuracy below.) For comparison, the corresponding definitions of TNF-alpha from the UMLS and OMD (Online Medical Dictionary) are given below:

TNF-alpha (UMLS) - *Serum glycoprotein produced by activated macrophages and other mammalian mononuclear leukocytes which has necrotizing activity against tumor cell lines and increases ability to reject tumor transplants.*
TNF-alpha (OMD) - *Originally described as a tumour inhibiting factor in the blood of animals exposed to bacterial lipopolysaccharide or Bacille Calmette Guerin.*

These examples highlight the fact that although the UMLS's and OMD's definitions are accurate, they may be too technical for the average user.

The goal of the DEFINDER project is to extract readable definitions and the terms they define. Automatic identification and extraction of terms from text has been widely studied in the computational linguistics literature [4] [5], and many systems exist for this task using both symbolic and statistical techniques [6]. The extraction of definitions and their associated terms has been less widely studied, although extraction of lexical knowledge has a rich literature [7] [8] [9].

We combine shallow natural language processing with deep grammatical analysis to extract definitions that are embedded in on-line full text. Through an analysis of a sample set of consumer-oriented articles, we identified typical cue-phrases and structural indicators that introduce definitions and the defined terms.

Our system is based on two main functional modules: 1) a pattern analysis module that performs shallow text processing using a finite state grammar, guided by cue-phrases ("is called", "is the term used to describe", "is defined as", "is the term for", etc.) and a limited set of text-markers ("()", "--"); and 2) a grammar analysis module that uses a rich, dependency-oriented lexicalist grammar for analyzing more complex linguistic phenomena (e.g. apposition, anaphora). The pattern analysis module is based on a surface part of speech tagger with a finite state grammar for identifying medical terminology and for extracting definitions. We used the Brill tagger [10] and the baseNP chunker [11] for identifying simple noun phrases (head noun + premodifiers). For the medical application, the lexicon was augmented with the most frequent medical terms found in our corpora. This eliminated incorrect tagging due to unknown words. A filtering module was added in order to remove some of the misleading patterns introduced by text markers (e.g. explanation, enumeration). The grammar analysis module is based on English Slot Grammar (ESG) [12]. The rich representation provided by ESG allows the identification of definitions that are introduced by more complex linguistic phenomena and not easily identifiable by shallow processing.

METHODS

In this section three methods for evaluating the DEFINDER system are presented: 1) performance in terms of precision and recall against a gold standard, 2) quality of extracted definitions in terms of user-based criteria of readability, usefulness and completeness, and 3) coverage of DEFINDER output vs. existing on-line dictionaries. For the first two, a user-centered evaluation using non-specialist subjects was performed. For the third we chose a set of defined terms extracted by our system and compared them to three on-line dictionaries [3]. The results 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.

DEFINDER Output vs. Gold Standard. The first evaluation method involves the comparison of DEFINDER output against a reliable "gold"

---
[3] a. The Merck Manual of Medical Information – Home Edition
b. Columbia University College of Physician & Surgeons Complete Home Medical Guide
c. Cardiovascular Institute of the South
d. Reuters Health Newspaper for Consumers
e. Medical Industry Today

standard. For this experiment we recruited four subjects who were not trained in the medical domain and who did not assist in the development of the system. Each subject was given a set of nine consumer-oriented articles chosen from our set of five text resources, and they were asked to manually mark-up definitions and their associated terms. We selected balanced examples from each genre (medical articles, newspapers, manual chapters, book chapters). Each subject was given instructions with examples of definitions found in articles similar to the ones we chose for the experiment. The gold standard consisted of 53 definitions identified by at least three out of four subjects. We measured DEFINDER performance in terms of precision and recall compared to this gold standard.

Quality of Definitions. We assume that non-technical definitions are more useful for consumers than specialized definitions. We evaluated the quality of DEFINDER output in comparison with two specialized on-line dictionaries (UMLS and OMD). Eight non-specialist subjects not connected with the project were provided with a list of 15 medical terms and with their definitions from each of the three resources. The source of each definition was not given in order not to bias the experiment. The task was to assign to each definition a quality rating for usefulness (U), readability (R) and completeness (C) on a scale of 1 to 7 (1 very poor, 7 excellent). Usefulness means that the definition could help the user to understand a technical term in the context of a technical article; readability means that the definition is easy to read and understand; completeness means that the definition is judged to contain full information about the term. As we discuss in the next sections, completeness is a less reliable feature for evaluation performed by non-specialist subjects but will be used in a future evaluation of accuracy by medical specialists.

We performed two studies for quality. In the first study, we measured the Average Quality Rating (AQR) for each of the three definitional sources on the three criteria. Our hypothesis was that DEFINDER would outperform both UMLS and OMD in terms of usefulness and readability, but that in terms of completeness, the on-line specialized dictionaries would be judged higher. We applied the sign test [13] to statistically validate our results.

One issue 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 this we performed a second study to evaluate the relative ranking of the three definitional sources (1 best, 3 worst). First we measured user agreement on ranking the definitions based on usefulness, readability and completeness on each individual term. Then for the terms for which the agreement was significant, we compute the overall mean ranks of the three sources and then measure again the significance of the results. In this analysis we used Kendall's coefficient of correlation, W [13], to measure interjudge reliability. The values of W are from 0 to 1, 0 showing no agreement, 1 showing perfect agreement. In computing W we include the corrections due to ties in ranking (when two or all three definitions of the term are given the same rank by one subject). A significant value for W means that the judges are applying essentially the same standard in ranking, thus eliminating the null hypothesis that the agreement is due to chance. As described in [13] the best estimate for the "true" ranking of the N objects (i.e., our three definition sources) is provided by the order of the mean ranks, given that W is significant.

Coverage of DEFINDER Output vs. On-line Dictionaries. The method for determining coverage is based on the comparison of DEFINDER output with existing on-line dictionaries. To quantitatively evaluate coverage, we selected three existing on-line dictionaries: the UMLS Metathesaurus, On-line Medical Dictionary (OMD) and Glossary of Popular and Technical Medical Terms (GPTMT). A base test set of 93 terms and their associated definitions, extracted by our system from text, was chosen for this experiment. Three cases were found: 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 any of the on-line dictionaries (absent).

RESULTS

DEFINDER output vs. Gold Standard. The resulting gold standard consisted of 53 definitions marked-up by at least 3 out of the 4 subjects. DEFINDER identified 40 out of these 53 definitions, obtaining 86.95% precision and 75.47% recall. Besides the correct definitions (40), DEFINDER extracted 6 false positive definitions, thus decreasing precision. An interpretation of these results is given in the Discussion section.

Quality of Definitions. In comparing the average quality rating values, results show that the definitions extracted by our system are judged higher than the definitions from the two other dictionaries in terms of usefulness and readability. In terms of completeness both UMLS and OMD show better results. Figure 1 shows the relative average quality rating (AQR) values of the three sources on all three characteristics: usefulness (**U**), readability (**R**) and completeness (**C**).

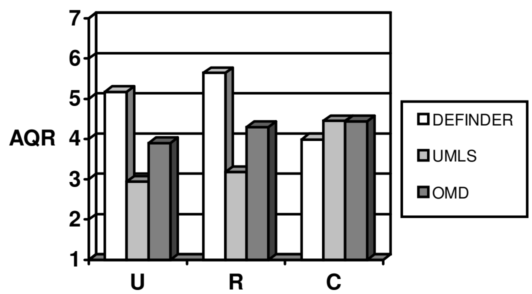

Figure 1 Average Quality Rating (AQR)

The sign test shows that the statistical significance of our results is p=0.0003 for usefulness and readability, and p=0.04 for completeness.

As previously mentioned, the results of the AQR can be noisy. In the second study we analyzed the relative ranking of the three definition sources, using Kendall's coefficient of correlation to statistically validate the results.

In terms of usefulness, users agreed in 13 out of 15 cases, with significant values of W ranging from 0.47 to 0.92 [13]. For these 13 terms we measured the level of correlation between them and then we computed the mean of ranks for the three definition sources. Since W=0.45 is significant, we can take as the "true" value of relative ranking the order provided by the mean ranks. Figure 2 shows that DEFINDER outperforms both UMLS and OMD in terms of usefulness.

In case of readability, the agreement between judges was significant in 14 out of 15 cases, W ranging from 0.56 to 1.00 (for 50% of the terms the values were above 0.85). The correlation between the rankings on these 14 terms was significant (W=0.54) and thus the ordering is relevant. As seen in Figure 2, DEFINDER definitions are judged higher on the readability scale than both UMLS and OMD.

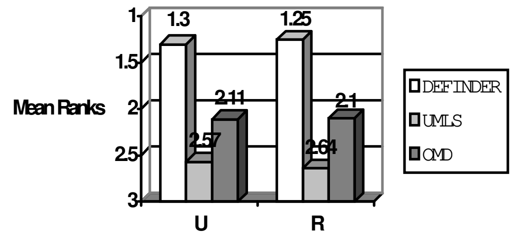

Figure 2 Ranking

Analyzing the agreement for completeness, in only 11 out of 15 cases was there significant agreement between subjects (W values between 0.38 and 0.92). The significance test for the correlation between the rankings of these 11 terms failed to provide a statistically representative value for W (W=0.26). Thus no inference regarding the order of rankings between the three definitional sources can be made. Reasons for this are presented in the discussion section below.

Coverage of DEFINDER Output vs. On-line Dictionaries. The results from Table 1 (also presented in [3]) show that on-line medical dictionaries are incomplete compared to potential DEFINDER output.

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

Table 1 Coverage of On-line Dictionaries

For example column three shows that in OMD only 71 definitions out of 93 possible definitions are found, thus giving only 76% completeness. GPTMT, although a glossary specifically addressed to lay users, is far from being complete; only 20 out of 93 terms were present, i.e. 21.5% coverage. This shows that DEFINDER identifies many terms and definitions that are lacking from existing resources.

DISCUSSION

The quantitative and qualitative evaluations presented in this paper show that the output of the DEFINDER system can be used in at least two ways: (1) to enhance on-line dictionaries and (2) to clarify terminology for non-specialist users by providing readable and useful definitions.

The results of the system performance in terms of precision and recall provide a strong baseline for further evaluating DEFINDER when applied on larger scale corpora. A careful analysis of human performance on identifying definitions and their associated terms from text shows that this is a very difficult task. Besides the 53 definitions, which constitute our gold standard, 8 definitions were identified by only one subject and 10 definitions by two subjects. The decrease in precision was because the system identified 6 false positive definitions. However, four out of these six definitions were also marked by one subject. The decrease in recall was because several definitions identified by human judges contain complex linguistic phenomena (anaphora or parallel definitions), not currently handled by our system. We expect these results to improve as we improve the DEFINDER system.

The qualitative evaluation of DEFINDER in terms of user-based criteria of usefulness, readability and completeness, shows that DEFINDER provides high quality definitions for non-specialist users. However our results raise a question regarding the evaluation of completeness and the ability of the lay user to rate on this criterion. In our future evaluation of DEFINDER on larger scale data, we will ask medical specialists to judge completeness and accuracy of the definitions.

The results on dictionary coverage showed that on-line dictionaries are incomplete. DEFINDER output can be used to address the gaps. In the UMLS, 24% of the terms were present but they belong to the axiomatic vocabulary, which in the case of specialized vocabularies is often highly technical and thus of limited use for lay people, e.g. “coumadin”, “Holter monitor”. Table 1 shows that 15 terms out of 93 were absent from UMLS. Following [14] we analyzed the missing terms and conclude that in some cases modifiers play an important role in deciding which are the “real” terms, e.g. “cardiac defibrillator” was the defined term extracted by our system, while in UMLS only the term “defibrillator” was present. Another example is “valvuloplasty” (DEFINDER) vs. “balloon valvuloplasty” (UMLS). Deciding which of these are terms is a task that requires specialist insight; we plan to further analyze these cases.

CONCLUSION AND FUTURE WORK

We present a definition finding system, DEFINDER that automatically mines online text for embedded definitions and their associated terms. We extract these definitions linked with the terms they define, and build a glossary. Our baseline results show 87% precision and 75% recall over a test set. Our results show high readability and usefulness of DEFINDER definitions, compared with existing on-line resources. In future work, we plan to: 1) extend to additional data; 2) address the issue of merging multiple definitions from different sources; 3) develop a method to evaluate for accuracy and completeness; and 4) integrate DEFINDER in the PERSIVAL medical information system.

REFERENCES

1. McKeown KR, Chang S-F, Cimino JJ, Starren J, Klavans, JL. PERSIVAL sytem. Available from: URL:http://www.cs.columbia.edu/diglib/Persival/
2. Hahn U., Romaker M. Text Structures in Medical text Processing: Empirical Evidence and a Text Understanding Prototype. Proceedings of AMIA Fall Symposium; 1997; p. 819-823.
3. Klavans JL., Muresan S. DEFINDER: Rule-Based Methods for the Extraction of Medical Terminology and their Associated Definitions from On-line Text. Proceedings of AMIA Symposium 2000; p. 1049.
4. Justeson J., Katz, S. Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text. Natural Language Engineering; Vol 1(1); 1995; p. 9-27.
5. Smadja F. Retrieving Collocations from Text:Xtract. In: Armstrong S, editors. Using Large Corpora. London: MIT Press; 1994; p 143-177.
6. Klavans JL., Resnik P, editors. The Balancing Act. Combining Symbolic and Statistical Aproaches to Language. The MIT Press; 1996.
7. Zeigenbaum 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.
8. Campbell DA, Johnson SB. A Technique for Semantic Classification of Unknown Words Using UMLS Resources. Proceedings of AMIA Annual Symposium; 1999; p. 716-720.
9. Johnson SB. Conceptual Graph Grammar – A simple Formalism for Sublanguage. Proceedings of IMIA WG6; 1997.
10. Brill E. A Simple Rule-based Part of Speech Tagger. Proceedings of the Third Conference on Applied Natural Language Processing. Trento, Italy; 1992
11. Ramshaw LA., Marcus MP. Text Chunking Using Transformation-Based Learning. Proceedings of Third ACL Workshop on Very Large Corpora, MIT; 1995
12. McCord MC. The Slot Grammar system. IBM Research Report; 1991
13. Siegal, S. and Castellan, NJ. Non-parametric statistics for the behavioural sciences 2nd Edition. New York: McGraw Hill; 1988.
14. McCray AT, Browne AC. Discovering the Modifiers in a Terminology Data Set. AMIA Annual Symposium; 1998 on CD-ROM [D004985]