A Method for Automatically Building and Evaluating Dictionary Resources

Smaranda Muresan*, Judith Klavans†

*Computer Science Department, Columbia University
500 West 120th, New York, USA
smara@cs.columbia.edu

†Center for Research on Information Access, Columbia University
535 West 114th St, New York, USA
klavans@cs.columbia.edu

Abstract

This paper describes a method toward automatically building dictionaries from text. We present DEFINDER, a rule-based system for extraction of definitions from on-line consumer-oriented medical articles. We provide an extensive evaluation on three dimensions: i) performance of the definition extraction technique in terms of precision and recall, ii) quality of the built dictionary as judged both by specialists and lay users, iii) coverage of existing on-line dictionaries. The corpus we used for the study is publicly available. A major contribution of the paper is the range of quantitative and qualitative evaluation methods.

1. Introduction

Most machine readable dictionaries or glossaries are either manually built by human experts or transformed in electronic forms from hard-copy versions through an expensive digitization process. Also for some particular domains, such as medical domain, the effort is concentrated in building *technical* dictionaries for specialists that are of little use for lay users who do not understand the jargon. Thus automatically building dictionaries as resources for natural language processing applications (e.g machine translation, summarization) would be a valuable new effort.

Definitions are by all means the most important part of any dictionary or glossary. In this paper we present a method for automatically extracting definitions and their headwords from text in order to build a new dictionary. Our current research focuses on medical domain, more particularly on consumer-health text.

Our contribution regarding resource building is twofold. One one hand, we construct a dictionary for non-specialist users. For example, the definition of the term *foam cells* extracted by our system, DEFINDER (Klavans and Muresan, 2000), from patient-oriented text is:

- *White blood cells that have ingested fat.*

while the technical definition present in UMLS (Unified Medical Language System) Metathesaurus [1]is:

- *Lipid-laden macrophages originating from monocytes or from smooth muscle cells.*

On the other hand, as shown in Section 3.4., on-line dictionaries are generally incomplete, so the output of our system can be used to fill in the gaps.

This research is part of Digital Library Project at Columbia University, entitled PERSIVAL (PErsonalized Retrieval and Summarization of Image, Video And Language resources) (McKeown et al., 2001). Our dictionary is used in the context of summarization of technical articles to provide explanation of the technical terms in lay language.

Our system, was extensively evaluated. First we evaluated the performance of the definition extraction method by comparing the results against a gold standard in terms of precision and recall. Second, we evaluated the quality of the dictionary as judged both by specialists and lay users. Third, we compared the coverage of several existing on-line dictionaries to see if our system can fill in the gaps.

Section 2 of the paper presents the corpora and the method for building automatically the lay medical dictionary. In Section 3 we describe both the quantitative and qualitative evaluations of our system. Section 4 presents our conclusions and future work.

2. DEFINDER - Using NLP Techniques for Automatic Dictionary Construction

Although there has been work on automatically inducing dictionary headwords (Schone and Jurafsky, 2001) and on term identification (Smadja, 1994; Justeson and Katz, 1995), the problem of automatically extracting definitions from corpora was less studied. Definitions are the most important part of any dictionary or glossary, thus automatically extracting/inducing them from corpora is a crucial piece in the effort of building dictionaries from text.

Automatic methods for text processing usually depend on the corpora on which they are applied. Thus in order to build an algorithm that will work in general settings, a variety of sources and text genre should be used.

2.1. Consumer-oriented Corpora for Lay Definition Extraction

In our research we are focusing on the extraction of definitions from consumer-oriented articles. The main sources of lay definitions are on-line consumer-oriented materials: articles, newspapers, book chapters, manuals, that are written by medical specialists in common language. Thus the main characteristic of lay definitions is that general language words are used to paraphrase the equivalent special-

[1]National Library of Medicine, http://umlsks.nlm.nih.gov/

ized terminology. Also depending on the context and on the author style, the definitions might not follow the *genus et diferentia* model (Byrd et al., 1987), thus making the work on machine readable dictionary parsing not suited for our task.

In order to make our algorithm work in general settings we select five corpora, of different genres: The Merck Manual of Medical Information - Home Edition, Columbia University College of Physician & Surgeons Complete Home Medical Guide, Cardiovascular Institute of the South (medical articles written by doctors for lay people), Reuters Health Newspaper for Consumers and Medical Industry Today.

A sample set of definitions of *myocardial infarction* from these articles is given below:

- *heart tissue death*
- *the most extreme state of oxygen deprivation, in which whole regions of heart muscle cells begin to die for lack of oxygen*
- *heart attack*

As can be seen, the style is different across definitions. In our approach we consider synonyms as definitions, which is a valid assumption in the theory of writing (Sager, 1990). Thus the variety of text genres and also the variety of authors writing styles pose a real challenge to computational techniques for automatic identification and extraction of definitions together with the headwords from full-text articles.

2.2. Automatic Method for Extracting Lay Definitions from Full-text

Like UMLS and other on-line medical dictionaries, we initially see the definitions as labels associated with terms, without semantic representation. In this light we developed a rule-based system, DEFINDER (Klavans and Muresan, 2000), that combines shallow natural language processing with deep grammatical analysis to identify and extract definitions and the terms they define from on-line consumer health literature.

DEFINDER is based on two main functional modules. The first module uses cue-phrases (e.g. *is the term for*, *is defined as*, *is called*) and text markers (e.g - -, *()*) in conjunction with a finite state grammar to extract definitions. We used the Brill's tagger (Brill, 1992) and the baseNP chunker (Ramshaw and Marcus, 1995) for identifying simple noun phrases. One problem is that the lexicon of Brill's tagger is derived from Penn Treebank tagset of the Wall Street Journal and Brown corpus. Thus unknown words in medical domain can cause serious errors in tagging. In order to alleviate this source of error, we augmented the lexicon with the most frequent medical terms found in our corpora. Relying on text markers might also provide a source of errors since they are used to introduce explanations and enumerations. Thus a filtering step is needed to eliminate the misleading patterns.

However this limited shallow analysis does not provide good recall when applied on the large variety of text genre and writing styles. To achieve higher accuracy, DEFINDER uses a grammar analysis module based on a statistical parser (Chaniak, 2000) in order to account for several linguistic phenomena used for definition writing (e.g appositions, relative clauses, anaphora).

The difference in writing style poses also the question of how we differentiate between the *headword* and the *definition*, in the case where both are simple noun phrases (i.e the definition is basically the lay synonym of the medical term). We used a simple statistical method based on frequency counts in order to differentiate. The hypothesis is that the *headword* is used several times in the article, while its definition tends to be mentioned only once.

3. Quantitative and Qualitative Evaluation

In order to thoroughly evaluate our system we extended our initial methodology (Klavans and Muresan, 2001), focusing on several dimensions: performance of the definition extraction algorithm in terms of precision and recall; quality of the generated dictionary as judged both by non-specialists and by medical specialists; coverage of on-line dictionaries.

3.1. Quantitative Evaluation of the Extraction Algorithm

A standard approach in any system evaluation is to compare the results against human performance. Thus we selected four subjects, not trained in the medical domain and who did not participate in the development of the system. Each of them was provided with a set of nine articles, and was asked to annotate the definitions and their headwords in text. We equally represent the sources of our corpora (medical articles, book chapters, manuals and newspapers), but we limit the length of the articles to two pages.

The gold-standard against which we compared our system was determined by the set of definitions marked up by at least 3 out of the 4 subjects and consists of 53 definitions. Our system obtained 86.95% precision and 75.47% recall.

The interpretation of the results was more difficult then expected, given that there was no agreement among users regarding *what is a definition?*, even though they were provided with a set of instructions and sample definitions. For example in the following sentence:

- The most frequent cause of the condition in older patients is atherosclerosis - the progressive narrowing of the heart's own arteries by cholesterol plaque buildups, which starves the heart itself for oxygen and nutrients.

our system identified as definition for *atherosclerosis*: *the progressive narrowing of the heart's own arteries by cholesterol plaque buildups,* ***which starves the heart itself for oxygen and nutrients***., while only 2 out of 4 subjects marked up the bold part.

3.2. Quality of the Built Dictionary - Lay User Perspective

As discussed in (Sager, 1990) an important aspect of the need for definitions is the user requirements. Satisfying both the specialist and the layman with a single definition

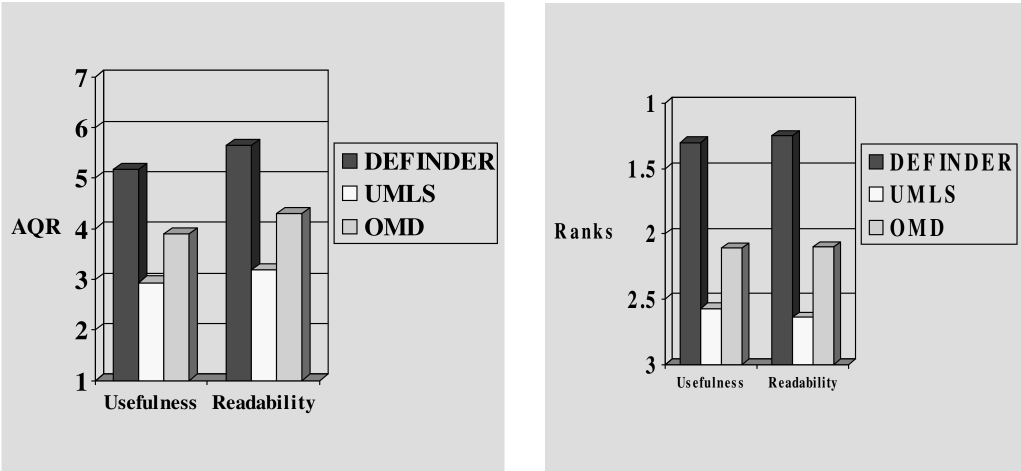

Figure 1: a) Average Quality Rating ; b) Mean Ranks

of a technical term will be hard to achieve. Thus, in our next evaluation, our aim was to compare the quality of our lay dictionary against existing specialized dictionaries from the perspective of non-specialist users.

We chose the UMLS Metathesaurus and the On-line Medical Dictionary (OMD) [2] as technical dictionaries. A set of eight subjects was provided with a list of randomly chosen 15 medical terms and their definitions from UMLS, OMD and the definition extracted by our system from on-line lay text. The source of each definition was not given in order not to bias the experiment. Also the order of definitions was randomly changed for each term. The task was to assign to each definition a quality rate (QR) for usefulness and readability on a scale of 1 to 7 (1 very poor, 7 excellent). Usefulness means that the definition can help the user understand the term, while readability means that the definition is not technical, thus is easy to read.

We first measured the Average Quality Rating (AQR) for each definition source on these two criteria. Our hypotheses were that DEFINDER outperforms both UMLS and OMD in terms of usefulness and readability. The results in Figure 1a support our claim. For usefulness, our system was rated 5.17 while UMLS and OMD obtained 2.94 and 3.9 respectively. In terms of readability, the difference was even higher: 5.65 compared with 3.18 and 4.3. In order to statistically validate our results we applied the sign test (Siegal and Castellan, 1988). As shown in Table 1 by the **p** values, the results were statistically significant.

Hypothesis	Usefulness	Readability
DEFINDER > UMLS	$p < 0.00003$	$p < 0.00003$
DEFINDER > UMLS	$p < 0.00003$	$p < 0.00005$

Table 1: Sign test (p) for Usefulness and Readability

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 subject, thus introducing noise in comparing the results. To address this issue we performed a second analysis to evaluate the relative ranking of the three definitional sources. Using Kendall's coefficient of correlation, W (Siegal and Castellan, 1988), we first measured the interjudge agreement 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 tested the same hypotheses: DEFINDER is better than UMLS and OMD both in terms of usefulness and readability. As Figure 1b shows DEFINDER indeed outperformed the specialized dictionaries. We obtained statistically significant W values (W=0.54 and W=0.45 at p=0.01 and p=0.05 respectively).

[2] http://www.graylab.ac.uk/omd

3.3. Quality of the Built Dictionary: Medical Specialist Perspective

The results of the previous section shows that the definitions extracted from consumer-oriented text are readable and useful for lay user, outperforming the existing specialized dictionaries. One question that arises is if they are also accurate and complete from medical point of view.

In order to answer this question we performed a user-based evaluation. We selected a set of 15 medical specialists (physician assistants, nurse practitioners, residents and medical students). Each subject was provided with the same set of 15 medical terms and the definitions extracted by DEFINDER from text, as the one given in Section 3.2.. They were asked to judge the accuracy and completeness of the definitions on a scale from 1 to 7 (1 very poor, 7 excellent).

The definitions were rated on average 5.87 for accuracy and 5.38 for completeness. The results show that consumer-oriented text, when of high quality can be a valuable source of definitions. Also because our definitions were embedded in text, one of their required characteristic was to be concise. This explain the somewhat lower value obtain for completeness.

3.4. Coverage of Existing Dictionaries

In this study we evaluated the coverage of three existent on-line dictionaries. In the introduction we claimed that these dictionaries are incomplete and our system can be used to fill in the gaps.

We selected two specialized dictionaries: UMLS Metathesaurus and On-line Medical Dictionary, and one popular glossary: Glossary of Popular and Technical Medical Terms (GPTMT) [3]. The popular glossary was chosen since it would be a good resource for lay users and we wanted to analyze its completeness. A base test set of 93 terms and their associated definitions was chosen for this experiment. As expected 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 one of the on-line dictionaries (*absent*)

Looking at the UMLS results, we noticed that 24% of terms were undefined, which is equivalent to say that they are in the axiomatic vocabulary. But the question is if these terms are really known by the lay users (e.g *Holter monitor* or *coumadin*)? Analyzing the terms that were classified as absent in UMLS, we conclude that modifiers play an important role in deciding which are the true terms (McCray and Browne, 1997) (e.g. *cardiac defibrillator* was the defined term extracted by our system, while in UMLS only the term *defibrillator* was present)

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

Table 2: Coverage of On-line Dictionaries

In the case of the popular dictionary(GPTMT) only 20 out of the 93 terms were present, thus achieving a coverage of only 21.5%. These results encourage us to believe that building automatically dictionaries from text is a valuable endeavor for enhancing existing resources.

4. Conclusions and Future Work

In this paper we described a method for automatically extracting definitions from text, as a key step in building dictionaries as resources for NLP applications. Our research focuses on medical domain, but the methodology of definition extraction can be applied to different domains.

Our method was applied on a corpus of consumer-oriented text in order to built a lay medical dictionary. The contribution of the work is twofold: on one hand, we provided an automatic method for dictionary construction that can be used for enhancing existing resources, and on the other hand, we provided an extensive methodology for evaluating our system. One future step is to apply a bootstrapping technique together with the rule-based method to increase the scalability of the system. Regarding the evaluation methodology we are planning to perform a usability evaluation of our dictionary in application context. We believe that our evaluation techniques are useful for the Computational Linguistics community.

Processing a vast amount of text of different genres, poses the challenge of extracting several definitions for the same medical term. The questions are: which definition to choose or how to merge all definitions into a more complex one? Our view is that there is not an *unique* definition suited for all applications (e.g in the context of summarization we may want a concise definition, while for enhancing dictionaries we might prefer a complex one). Our future research work will address these issues.

[3] http://allserv.rug.ac.be/%7Ervdstich/eugloss/welcome.html

5. References

E. Brill. 1992. A simple rule-based part of speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*.

R. J. Byrd, N. Calzolari, M. Chodorow, J. Klavans, M. S. Neff, and O. A. Rizk. 1987. Tools and methods for computational lexicology. *Computational Linguistics*, 13(3-4):219–240.

E. Chaniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of NAACL-2000*.

J. Justeson and S. Katz. 1995. Technical terminology: Some linguistic properties and an algorithm for identification in text. In *Natural Language Engineering*, volume 1, pages 9–27.

J. Klavans and S. Muresan. 2000. DEFINDER: Rule-based methods for the extraction of medical terminology and their associated definitions from on-line text. In *Proceedings of AMIA Symposium 2000*.

J. Klavans and S. Muresan. 2001. Evaluation of DEFINDER: A system to mine definitions from consumer-oriented medical text. In *Proceedings of The First ACM+IEEE JCDL 2001*.

A.T. McCray and A.C Browne. 1997. Discovering the modifiers in a terminology data set. In *AMIA Annual Symposium*.

K. McKeown, S-F Chang, J.J Cimino, and et al. 2001. PERSIVAL, a system for personalized search and summarization over multimedia healthcare information. In *Proceedings of The First ACM+IEEE JCDL 2001*.

L. Ramshaw and M. Marcus. 1995. Text chunking using transformation-based learning. In *Proceedings of Third ACL Workshop on Very Large Corpora*.

J.C. Sager. 1990. *A Pratical Course in Terminology Processing*. John Benjamins Publishing Co. Amsterdam.

P. Schone and D. Jurafsky. 2001. Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*.

S. Siegal and N.J. Castellan. 1988. *Non-paprametric statistics for the behavioural sciences*. New York, McGraw Hill, 2nd edition.

F. Smadja. 1994. Retrieving collocations from text:xtract. In S. Armstrong, editor, *Using Large Corpora*, pages 143–177. London: MIT Press.