Improving Content Based Recovery on a Radiological Reports Database

Fábio Alexandrini[1], Mariana Kessler Bortoluzzi[2], Aldo von Wangenheim[3]

[1,2,3] The Cyclops Project Department of Computer Science Federal University of Santa Catarina: 88049-900 Florianópolis, SC, Brazil
[2] Department of Business Information Systems II University of Trier 54286 Trier, Germany
[1]fabalex@unidavi.edu.br; [2]kesslerb@uni-trier.de; [3]awangenh@inf.ufsc.br.

Abstract. The present effort focuses on developing a method for assisting the representation of radiological reports, written in simplified natural language, in a standardized and content recovery prone structure, such as DICOM Structured Report. Sample reports were collected and have being analyzed. The work is currently in process, but an intermediary representation was already reached and is being evaluated by physicians to attest the accuracy of the results.

1 Introduction

Most health care institutions have a precious legacy of clinical reports written in natural language, or simplified grammatical structure. Unfortunately content based retrieval of information from these reports is inefficient due to the peculiarities of natural language. This prevents institutions from sharing clinical records without waste of precious time and resources.

The present research effort focuses on the development of methods based on knowledge about normal findings in radiological examinations [1] and the Systematized Nomenclature of Medicine - SNOMED [2] with the objective of translating thoracic radiological reports into a representation more suitable for content recovery and that can be, in further work, rendered into reports compliant to internationally accepted standards such as DICOM Structured Report [3]. A set of radiological reports, provided by a Brazilian and a German health care institution, was used as source of sample subjects for the experiment.

2 State of the art

The natural language processing involves the development of intelligent computer systems that deals with problems in microworlds, application limited domains characterized by the search of most appropriated technique to solve each sort of problem. It seeks to develop systems capable to solve complex problems composed by distinguished tasks [4]. Some subfields of infomation retieval rely on a training corpus of documents that have been classified as either relevant or non-relevant to a

particular situation, in text categorization or attempts to assign documents to two or more pre-defined categories [5].

In the same way there must be a more appropriated technique to solve each task, where several tools of AI and other areas of knowledge are combined in only one intelligent system that manages the efforts to solve tasks.

Clinical report, especially in radiology contains information concerning a patient's medical condition. However, a great percentage of this information is not structured, for it's free text based, that consequently makes it more difficult to search, analyze, summarize and present.

Previous studies have shown the potential benefits of medical structured data for the practice, research and medical teaching. The information's structure can be used to help organize and improve the medical record presentation [6], [7], [8], [9], [10]. Specialist systems can use the structured information of Clinical report to decision support [11], [12], [13]. For research and teaching, structured Clinical Reports can extremely improve recall and the precision of recovering information's tasks. Only structured data are accessible to the cause, space, time and evolutionary advanced database that models the technique being developed on computing and medical informatics fields [14], [15], [16].

Other systems that seek to enlarge the semantic contents of ontology to guide the knowledge discovery on database and also analyze the variables types that the user checks [17]. Systems that work on evaluation and comparison of terminology models use of diagnosis concepts for clinical terms terminology for integration, as SNOMED, have obtained success with semantics categories of ECS – European Committee for Standardization of structured categories and ISO reference model of terminology [18]. But there are also problems concerning information quality, one of the main factors of successfully or unsuccessfully cases of application or methodologies proposed.

One fact that deserves to be singled out is the lack of works directed to applications in the Portuguese Language, face to the fact that most terminologies are found mostly in English, having a few versions to other languages as German, French and Spanish. Portuguese is the eighth most spoken language on the planet, third among occidental languages, after English and Spanish [19].

3 The DICOM SR Standard and the SNOMED Nomenclature

SNOMED is a widely accepted terminology and infrastructure designed to enable the sharing of health care knowledge, across clinical specialties and sites of care. It contains, preferred medical terms and concepts, consisting of more than 144,000 terms and term codes divided into eleven: Topography; Morphology; Function; Living Organisms; Chemicals, Drugs, and Biological Products; Physical Agents, Activities and Forces; Social Context; Diseases/Diagnoses; Procedures; and general Linkage Modifiers. SNOMED is available in several languages including German. Unfortunately there is not yet an available translation of the SNOMED terms for Portuguese. Thus, in order to make this experiment possible, the basic terms for thoracic radiological exams were translated by collaborator Brazilian physicians.

The DICOM SR standard sets out rules that define how structured documents that contain health information should be composed, stored and transmitted. These make use of a controlled terminology; which enhances the results of content based retrieval [20].

4 Methods

The sample thoracic reports, 315 of them in German and 7719 in Portuguese, were analyzed. In the sample reports a common structure was identified. A heading containing the identification information such as of the physician, patient, and institution in which the report was emitted, followed by a description of the techniques used to perform the exam, such as type of procedure. The main part of the report is the body in which the conclusions drawn from the exam can be found. This will be the focus of the effort to interpret, retrieve and codify content.

The language used by physicians to describe findings in radiological examinations in both the institutions can not be considered indeed natural language. It is rather a set of short sentences normally without the use of verb, escaping the linguistic conventional rules. These, often either affirm of deny the existence of a malformation or alteration in an anatomical structure, the presence of strange bodies or illnesses, adding observations regarding the localization, shape or appearance of the anatomical structure. For each type of clinical report of radiological examination there are specific anatomical structures of interest [1].

The subject of most sentences found in thorax radiological reports are anatomical structures such as *mediastinum*, *lung*, *trachea* and others. The subject is often followed by adjectives stating the position, or the part of the organ that was observed, for instance *surface* or *right*. Adverbs, describing morphological information follow, such as *normal* or *reduced*. To this, might follow information regarding diseases, like for example, the existence of emphysema and adverbs indicating the degree of severity of the disease. For the analysis and structuring process ware used natural language processing techniques, lexical, syntactic, semantics and of speech analysis combined with information from the specialist physicians. Because the medical language and usually language have a great difference, normality the language in the clinical reports are affirmative or negative phrases with few or without verb.

5 Structuring Approach

The sample radiological reports from Dataflex and word documents base were converted to simple text ASCII standard representation in order to disregard formatting information. Afterwards, the GNU Aspell free tool [21], available in several languages, was used to perform a lexical checking.

For the natural language processing, the text file is separated in sentences observing punctuation and end of line commands. Later the words are separated for the lexical analysis. Using a former developed dictionary, the words are classified by their grammatical class, such as substantive, adjective and others. The sentences and

words are the input for the syntactic and semantics analysis. Each word has a function in the sentence, but we search first the terms for anatomical structures, diseases and morphology are then matched to the appropriate SNOMED codes.

After the first tests of classification, there was found a great amount of adjectives that derived from substantives and for that reason had a huge importance, needing then a special treatment, for were directly referred to anatomies. See the table 1. For each sentence without an anatomical SNOMED matching term, a speech analysis is used, the algorithm searches an anatomical term from the previous phrases. When no term is found, the code of anatomical term from the clinical report's title is used.

Table 1. Example of word classification

Word			grammatical class	Snomed	Refe-rence
Portuguese	**English**	**Germany**			
Pulmão	Lung	Lunge	substantive	T28000	
pulmonar	pulmonary	pulmonal	adjective	T28000	Lung
Pleura	Pleura	Brustfell	substantive	T29000	
pleural	pleural	pleural	adjective	T29000	Pleura
normal	normal	normal	adjective		
Alteração	Alteration	Änderung	substantive		
anatômico	anatomic	anatomisch	adjective		

Earlier the use of specific rules for each type of exam was used, but due to the complexity and difficulty to make a great number of rules, computationally were studied the resemblance and differences among the medical reports and the sentence structures contained on those. Despite the diversity of sentences as well as in numbers of words as in the way of position or construction, the most important information to be extracted are the anatomical region (where?), the diseases or alterations (what?), and the qualifications of those (how?). Additionally can be used the position into anatomical region and general modifiers, such as *right*, *left*, *bilateral*, *surface*, *superficial* and others.

The terms for anatomic regions are searched in the "Topography" branch of SNOMED. If a word regarding position is along with the anatomic region, that also can be converted to one of the terms in the "General Linkage and Modifiers" branch of the SNOMED hierarchy, and the diseases or alterations (what?) can be converted in categories "Morphology", "Diseases/Diagnoses" or "Function". In this part of the process we obtained the support from the Teachers and Resident Doctors in General Surgery from HRAV (Hospital Regional Alto Vale).

In association with the overall are found the terms of qualification (how?) such as: *normal*, *intact*, *anatomic*, etc, and can also be combined with the denial as: *without*, *absent*, etc, that indicate the absence of problems, or their intensity such as: *minimum*, *light*, *acute*, *chronic*, etc. In the following example of analysis, on the first original sentence of the chart, there are three anatomic regions and one qualification, that

would generate three sentences according to the where and how items, in a relation N anatomies to 1 qualifier, and the word "*bronchi*" also applies to "*lobar*".

On the second sentence is noticed the relation of 1 to 1 and on the last sentence is found a relation of 1 anatomy to N qualifiers, besides finding the pleural derived object that needs to be converted for the substantive *Pleura* added by the adjective *surface*, that can be found respectively in the "Topography" and "General Linkage Modifiers" branches of SNOMED. For each anatomical SNOMED term and adjective in a phrase, a new sentence must be created; this process can be visualized on the Table 2.

Table 2. Example of sentence analysis

Original Sentence	**Where**		**How**	Sentences Generated
	Anatomy (Substantives)	**Snomed**	**Qualification (Adjectives)**	
Trachea, Main and Lobar Bronchi, of normal gauge.	Trachea	T25000	Gauge Normal	3
	Main Bronchi	T26000		
	Lobar Bronchi	T26200		
Pulmonary Anatomic Hilo	Pulmonary Hilo	T28080	Anatomic	1
Pleural anatomic surface without alteration	Pleura Surface = **Position**	T29000 G-A168	Anatomic	2

6 Conclusion

The language used by physicians to describe findings in radiological examinations in both collaborator hospitals is a simplified restricted one. Nevertheless, there is much research still to be done before a reliable representation of such reports in a standard such as DICOM SR can be obtained.

The objective of this work is not to produce software capable of performing a complete automatic translation of clinical reports written in natural language to a standardized form, but to develop an approach to facilitate de process of structuring documents, so that these are better suited for content based retrieval.

Although it is currently being tested using reports written in Portuguese and German, the same method slightly adapted is expected to function with other languages. The results so far reached are at the present time subject to the evaluation of the physicians to attest the validity and suggest better approaches.

Acknowledgements

The authors thank the physicians of the Chirurgic Residence of the Upper Itajaí Valley Regional Hospital in Brazil for the translation of Radiological SNOMED terms to Portuguese and for the anonymized sample reports provided for the analysis. The Buddenbrock Blasinger und Benz Radiological Clinic, in Mainz, Germany, contributed providing anonymized sample reports. We also thank the German Academic Exchange Service -DAAD- for the scholarship number A/03/42304 granted to one of the authors.

References

1. Möller, Torsten B.: Normal Findings in Radiology. Georg Thieme Verlag, 2000
2. College of American Pathologists: SNOMED - Systematized Nomenclature of Medicine. College of American Pathologists, 1994
3. NEMA.: Digital Imaging and Communications in Medicine (DICOM): Version 3.0; 2000
4. Russel, S., Norvig, P., Artificial Intelligence - A Modern Approach. Pearson Education Inc, 1995
5. Manning, C., Schütze, H.: Foundations of Statistical Natural Language Processing 6.ed. MIT. Massachusetts, 2003
6. Langlotz, Curtis P., Automatic Structuring of Radiology Reports: Harbinger of a Second Information Revolution in Radiology, Radiology, 2002
7. Hripcsak, G. et al.: Use of Natural Language Processing to Translate Clinical Information from a Database of 889,921 Chest Radiographic Reports. Radiology, 2002
8. Ricky, K. T., Stephen, G., Soderland, R. M. J.: Automatic Structuring of Radiology Free-Text Reports. Radiology, 2001.
9. Shortliffe, E.H., Hubbard, S.M.: Information systems in oncology. In: De Vita VT, Hellman S, Rosenberg S, eds. Cancer: principles and practice of oncology. Philadelphia, Pa: Lippincott, 1989
10. Aberle, D.R., et al.: Integrated multimedia timeline of medical images and data for thoracic oncology patients. RadioGraphics, 1996
11. Lyman, M, S. N., et al.: The application of natural-language processing to healthcare quality assessment. Med Decis Making, 1991
12. Brown, P. J. B., Warmington, V.: Data quality probes—exploiting and improving the quality of electronic patient record data and patient care. International Journal of Medical Informatics, 2002
13. Sager, N., et al.: Medical language processing: applications to patient data representation and automatic encoding. Methods Inf Med, 34:140–146, 1995
14. Muller, R., et al.: A graph-grammar approach to represent causal, temporal and other contexts in an oncological patient record. Methods Inf Med, 35:127–141, 1996
15. Abidi. R., Manickan, S.: Extracting Case Structures from XML-Based Electronic Patient Records: A Knowledge Engineering Solution to Augment Case Based Reasoning Systems. International Journal of Medical Informatics, 2002
16. Jackson, P.: Natural Processing for online applications. J.Benjamins Pulishing Co., Philadelphia, 2002
17. Phillips, J., Buchanan, B.G.: Ontology-guided knowledge discovery in databases. International Conf. Knowledge Capture Victoria, Canada, 2001
18. Bakken, S., Warren, J. J.: An evaluation of the usefulness of two terminology models for integrating nursing diagnosis concepts into SNOMED Clinical Terms ®, International Journal of Medical Informatics, 2002

19. Medeiros, A.: A Língua Portuguesa [On-Line] available online URL: http://www.linguaportuguesa.ufrn.br, 2004
20. Clunie, David A.: DICOM Structured Reporting. PixelMed Publishing, 2000
21. Atkinson, K.: GNU Aspell. Available online URL: http://aspell.net/, 2004