Information Extraction from Mammogram Reports

Anna Kupść,
Małgorzata Marciniak,
Agnieszka Mykowiecka
IPI PAN
Ordona 21
01-237 Warszawa,
Poland
{aniak,mm,agn}@ipipan.waw.pl

Jakub Piskorski
DFKI GmbH
Stuhlsatzenhauseweg 3
D-66123 Saarbuecken,
Germany
piskorsk@dfki.de

Teresa Podsiadły-Marczykowska
IBIB
Trojdena 4
02-109 Warszawa,
Poland,
teresa@ibib.waw.pl

Abstract

In this paper, we present an environment designed for extraction of medical data from mammogram reports. We process data collected from various Polish health care providers and transform them into attribute-value structures, according to a simplified mammographic ontology. We use a general purpose information extraction (IE) platform, SProUT, enriched with domain-specific terms. We adopt a cascaded processing strategy and merge externally the results obtained by IE techniques. To the best of our knowledge, the current project is the first attempt at IE from Polish medical texts.

1 Introduction

The past few years have witnessed a growing interest in applying NLP techniques to process and understand biological and medical texts. There have been created many resources and processing tools which facilitate access to desired information. However, most of these resources are monolingual and cannot be directly reused for other languages. In this paper, we present the first attempt at automatically obtaining structured information from Polish medical texts. A similar task for English mammograms was undertaken by (Hahn et al., 2002). (Burnside et al., 2000a) used a Bayesian network to identify finding's features, while (Burnside et al., 2000b) proposed a statistical method for mapping radiology reports to BI-RADS (*Breast Imaging Reporting and Data System*) terms.

The aim of the project described in this paper is to provide a formalized description of Polish mammogram reports. As a starting point, we take a detailed hand-crafted ontology. However, to make our task realistic, we build a simplified domain model adjusted to our needs. Then, we extract partial information from the texts using SProUT (Drożdżyński et al., 2004), a general-purpose IE platform, which has been adapted to processing of Polish (Piskorski et al., 2004). Finally, we combine extracted phrases together so that separate pieces of information fit our domain model.

2 Mammographic Ontology

The Polish mammographic ontology, a conceptual model of restricted subfield of radiology — mammography, has been based on medical literature (D'Orsi and Kopans, 1993), (Kopans et al., 1993), (Dziukowa, 1998), interviews with expert radiologist and knowledge found in the corpus (around 3000 routine free-text mammogram reports). During ontology development, concepts relevant for the domain, their properties and relations between them have been identified, domain-specific terms (with synonyms), referring to concepts and attributes have been collected. Properties of concepts have been separated into two groups: visual features of mammographic findings describing their appearance on the film, and non-visual features such as radiological diagnosis, assessment, subjective interpretation or recommendations.

The main root class in the mammography model is the concept of Mammographic Observation. Its direct subclasses are: *Breast Composition, Breast Finding and Axillary Lymph Node*. Instances of these classes can be used to create a knowledge base of mammography, containing important mammographic lesions described in the literature. At the moment there are 130 classes, 342 slots and 58 instances in the model. The model has been formalized using Protégé-2000 (`http://protege.stanford.edu/`), a frame-based ontology editor.

In order to make our task realistic, we have

simplified the above model and adjusted it to our needs. The simplified model is represented by attribute-value pairs (AVMs). An attribute's value can be atomic, another AVM, a list of atomic values or AVMs (limited by $< >$). A structure defined for representing a report is sketched in (1).

(1)
$$\begin{bmatrix} \text{EXAM\_ID exam\_id\_value} \\ \text{EXAM\_DATE exam\_date\_value} \\ \text{EXAM\_ID pat\_id\_value} \\ \text{R\_B brest\_desc\_t} \\ \text{L\_B brest\_desc\_t} \\ \text{FINDINGS} < \begin{bmatrix} \text{CHANGE change\_t} \\ \text{LOC loc\_t} \end{bmatrix} > \\ \text{R\_LYMPH\_NODES } \textit{nodes\_desc\_t} \\ \text{L\_LYMPH\_NODES } \textit{nodes\_desc\_t} \\ \text{DIAGNOSIS diagnosis\_t} \\ \text{RECOMMEND recommend\_t} \end{bmatrix}$$

The first group of attributes contains general information about the examination, e.g., an identification number, examination date or a patient identification number.

The next few attributes describe breasts' composition. type of tissue with detailed information about glandular tissue, its localization, density, regularity, a comparison with the previous examination, and recommendations resulting from the above data.

Another attribute is FINDINGS. Its value is a list of AVMs representing the findings encountered in the mammogram report. Each finding is described separately by the following attributes: ANAT_CHANGE — the finding's appearance on the film, e.g., darkness, tumor or tissue concentration; LOCALIZATION — an AVM specifying an anatomic localization, a body part, lateralization and a conventional localization; DENSITY; SHAPE; MULTIPLICITY represents both the exact (*cztery* 'four') as well as an approximate (*nieliczne* 'a few') amount of identified findings; CONTOUR; SIZE — an AVM with attributes describing up to three dimensions and a measurement unit; WITH_CALCIF — information about accompanying calcifications (micro or macro); PALPABILITY — a 'yes' or 'no' value ; INTERPRETATION of the finding (a cyst, cancer, an intramammary lymph node, etc.); DIAGNOSIS_RTG — information wheather the finding seems to be benign, suspicious or malignant; RECOMMEND — further examinations required and CHANGES_IN_TIME — the finding's changes in time.

The next group of attributes contains information about lymph nodes as well as their description and diagnosis. The report often ends with a general recommendation or diagnosis. So we have the DIAGNOSIS and RECOMMEND attributes at the main level of the AVM structure.

3 System Architecture

In the project, we adopted a cascaded processing strategy and divided the extraction process into four stages: pre-processing, basic IE, cleaning-up of the extracted data, and final merging.

The pre-processing stage was motivated by the low quality of texts produced by physicians. There are many spelling errors (mostly lack of Polish diacritics but also other misspellings) and punctuation errors (lack of commas, periods or their non-standard usage) as well as domain-specific abbreviations. Hence, using uncorrected data would result in a severe data loss.

Polish is a language with rich inflection, so extraction from Polish medical text requires not only recognition of medical terms but also their inflected forms. Unfortunately, we have no access to a Polish electronic medical lexicon. On the other hand, many medical terms are present in everyday speech and are covered by general-purpose dictionaries. We used a general morphological analyser integrated with SProUT (Piskorski et al., 2004), which also allowed us to employ SProUT rules to recognize more complicated syntactic forms, not just isolated words. In particular, SProUT enables building phrases on the basis of morphological features of their elements. Since only part of the medical terminology can be recognized (and inflected) by the morphological analyser, we also employ a specialized lexicon integrated with SProUT — the so-called gazetteer — which can store unrecognized mono- and multiword expressions.

IE from unstructured texts requires a tradeoff between simplicity and extraction completeness. This becomes extremely important when dealing with data corresponding to related features appearing freely in the text. For example, this happens when we collect all information about a particular finding (its shape, size, contour, density, change in time, localization etc.). In addition to various permutations, pieces of relevant information can be scattered in the document

775 W sutku prawym przybrodawkowo widoczny guzek o śr. 10mm z makrozwapnieniami w jego obrębie odpowiadający f-a degenerativa (zmiana łagodna). W sutku lewym w KGZ wewnątrzsutkowy węzeł chłonny.
[In the right breast in subareolal, there is a tumor of 10mm diameter with calcifications corresponding to f-a degenerativa (bengin finding). In the left breast, there is an intramammary lymph node in the upper outer quadrant.]

Figure 1: Sample mammogram report

and it will be impossible to merge them locally. Therefore, we process data sequentially. First, we use SProUT to identify all pieces of relevant information. Then, the results are externally processed and merged into complex AVMs. This process is divided into two steps: first, we clean up the data from unneccesary information, and then we group search results into blocks, according to our domain model presented in section 2.

4 Merging

After cleaning up the extraction results, a sequence of attribue-value pairs which specify the recognized phrases, e.g., `EXAM_ID:775`, is stored (each on a separate line) in a text file. The last processing phase identifies blocks corresponding to a finding description.

The two main types of blocks represent findings and the breast's composition and are marked `zp` (`zk`) — start (end) of a finding description, and `up` (`uk`) — start (end) of the breast's composition description.

The annotation process is based on the position of attributes representative for each type of block, i.e., ANAT_CHANGE, INTERPRETATION (for findings), and BTISSUE (for breast's composition). Lines containing these attributes are tagged, respectively, `a_ch`, `i_ch` and `ut`. All lines with attributes which do not belong to any block (e.g., DIAGNOSIS_RTG_LOC or attributes starting with BR_) are marked as `dloc`. The last part of the report, containing general RECOMMENDATIONS, is marked with the `rp` tag. The process of identifying blocks is repeated, starting from the first line marked with `a_ch`, `i_ch` or `ut` tags. From that line we go back to the previous block's opening or closing tag, and then go forward, trying to cover the maximal part of the report unless the `dloc` tag or attributes unique for a finding (e.g., localization, shape, size) are found. In this case, the corresponding closing tag (`uk` or `zk`) is inserted.

```
-- EXAM_ID:775
zp
-- LOC|BODY_PART:sutek
 ||LOC|LOC_CONV:ok. brodawki sutkowej||LOC|L_R:prawy
-- ANAT_CHANGE:guzek||MULT:singular
-- DIM:mm||NUM1:10||NUM2:10
-- C_MULT:plural ||WITH_CALCIF:makrozwapnienie
-- INTERPRETATION:f-a degenerativa
-- DIAGNOSIS_RTG:zmiana_lagodna
zk
zp
-- LOC|BODY_PART:sutek
 ||LOC|LOC_CONV:loc_KGZ||LOC|L_R:lewy
-- INTERPRETATION: wewnątrzsutkowy węzeł chł.
zk
```

Figure 2: Processing result for example in Fig. 1

123 Sutek prawy – w kwadrancie górnym zagęszczenie dobrze wysycone o średnicy około 20 mm i zatartych granicach. Wymaga ona dalszej diagnostyki – konieczne wykonanie badania USG i PCI. Wewnątrzsutkowy węzeł chłonny w kwadrancie górno-zewnętrznym sutka lewego.
[The right breast – in the upper outer quadrant there is a high density finding ofabout 20 mm diameter and obscured margins. Requires further examination – USG and biopsy compulsory. An intramammary lymph node in the upper outer quadrant.]

Figure 3: Sample mammogram report

Sample processing results for reports in Fig. 1 and 3 are presented, respectively, in Fig. 2 and 4. In Fig. 2, identifying a new localization (unique for a finding) is a good criterion for separating findings' descriptions. In some reports however this strategy leads to wrong segmentations. In Fig. 3, for the second finding, only the interpretation is given. As its localization occurrs after this finding and there is no interpretation for the first one, 'intramammary lymph node' is classified as an interpretation of 'density'.

The evaluation of the presented method is provided in Fig. 5.

We have identified the following main reasons of detected errors: 1) coordination — some ele-

```
-- EXAM_ID:123
zp
-- LOC|BODY_PART:sutek
 ||LOC|LOC_CONV:loc_KG||LOC|L_R:prawy
-- ANAT_CHANGE:zagęszczenie ||MULT:singular
-- SATURATION:dobrze wysycony
-- DIM:mm||NUM1:20||NUM2:20
-- CONTOUR:zatrzeć zarysy
-- RECOMMENDATION:USG_PCI||TIME:unknown
-- INTERPRETATION: Wewnątrzsutkowy węzeł chł.
zk
-- LOC|BODY_PART:sutek
 ||LOC|LOC_CONV:loc_KGZ||LOC|L_R:lewy
```

Figure 4: Processing result for example in Fig. 3

	nb	%
NUMBER OF PATIENT RECORDS	448	
FINDINGS	474	100
correctly recognized beginning of findings	416	87,76
unrecognized findings	13	2,74
incorrectly recognized findings	17	3,59
findings with an incorrect beginning	45	9,50
SAMPLE ATTRIBUTES		
SATURATION	185	100
correctly recognized	182	98,38
WITH_CALCIF	40	100
correctly recognized	35	87,50

Figure 5: Evaluation of findings' descriptions

ments of conjoined phrases are not repeated and in most cases this results in identifying only one of the conjoined elements; 2) negated phrases — not all forms of negation have been captured by shallow extraction rules, which caused opposite interpretations; 3) paraphrases — different ways of expressing the same concept disallowed its full recognition.

5 Conclusions and Future Work

The paper presents a combined approach to IE from mammogram reports. IE from brief and compacted texts, meant originally as notes for other physicians, turned out to be a quite challenging task. Main processing issues were caused by the lack of a clear document structure, style differences between various physicians, paraphrases, laconic formulations and the intensive use of idiosyncratic abbreviations. Another problem was a discrepancy between the general mammographic ontology, developed mainly on the basis of medical knowledge, and formulations found in the reports: very often they could not be directly translated into ontology concepts as statements used in reports were unclear, incomplete or ambiguous.

We divided the extraction process into several steps. As the information we needed to extract was often scattered in the reports, we decided to first extract smaller pieces of information and, then, combine them externally into more complex AVMs. This solution turned out to be quite successful in collecting the data but still a lot of problems have to be resolved. For example, elliptic references to previous findings, as in 'There are several changes in the left breast, *the greatest* of 2cm size' or relative phrases such as '*a similar* finding' remain uninterpreted. Hence, the next step is to enhance the grouping method so that more complex cases are covered.

We also plan to incorporate an inference mechanism. This would allow for filling in data missing from the reports but which can be inferred based on general medical knowledge. After the amendments, data will be entered to a database where they can be further analysed. We intend to take a full advantage of SProUT's option of defining a cascade of IE-grammars.

References

E. Burnside, D. Rubin, and R. Shachter. 2000a. A Bayesian Network for Mammography. In *Proceedings of the American Medical Informatics Association Symposium*, pages 16–110.

E. Burnside, H. Strasberg, and D. Rabin. 2000b. Automated Indexinng of Mammography Using Linear Least Squares Fit. In *CARS 2000 International Conference on Computer Assisted Radiology and Surgery, San Francisco.*

C. J. D'Orsi and D.B. Kopans. 1993. Mammographic Feature Analysis. *Seminars in Roentgenology*, 28:204–230.

W. Drożdżyński, H-U. Krieger, J. Piskorski, U. Schäfer, and F. Xu. 2004. Shallow Processing with Unification and Typed Feature Structures — Foundations and Applications. *German AI Journal KI-Zeitschrift*, 01/04. Gesellschaft für Informatik e.V.

J. Dziukowa. 1998. *Mammografia w Diagnostyce Raka Sutka.* PWN, Warszawa.

U. Hahn, M. Romacker, and S. Schultz. 2002. MEDSYNDIKATE — a natural language system for the extraction of medical information from findings reports. *International Journal of Medical Informatics*, pages 63–74.

D. B. Kopans, C.J. D'Orsi, D.D Adler, and al. 1993. Breast Imaging Reporting and Data System (BI-RADS). In *American College of Radiology.*

J. Piskorski, P. Homola, M. Marciniak, A. Mykowiecka, A. Przepiórkowski, and M. Woliński. 2004. Information Extraction for Polish Using the SProUT Platform. In *Intelligent Information Processing and Web Mining. Proceedings of the IIS'04 Conference, Zakopane.* Springer.