

# Rule-based Medical Content Extraction and Classification

Agnieszka Mykowiecka, Anna Kupść, Małgorzata Marciniak

Institute of Computer Science, Polish Academy of Sciences,  
Ordona 21, 01-237 Warsaw, Poland

**Abstract.** We present the final version of the system for automatic content extraction from Polish medical data. The system combines general IE techniques with an external post-processing. The obtained data is normalized and linked to a simplified ontology. Then, it is automatically grouped to form more complex structures representing medical reports.

## 1 Introduction

The paper describes the final version of the system used for automatic content extraction from Polish mammogram reports. The system serves to extract and standardize mammographic data so that it can be stored in a uniform form in a database which will support physicians in decision making and diagnosing.

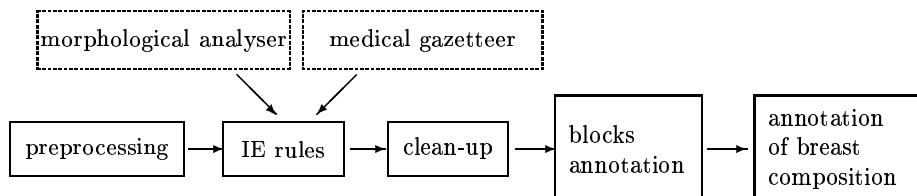
The extraction process is divided into template extraction using SProUT (a general-purpose IE system, adapted to Polish, see [7]) and external merging of the extracted data into more complex structures, according to our domain model presented in [5]. Three main types of blocks have been identified in the reports: findings, breasts' composition and the overall diagnosis. The merging procedure has been specialized for each part of the report, as different types of information require a separate treatment. The extracted data is normalized and linked to a simplified mammographic ontology. The ontology is based on the general domain knowledge and the analysis of medical reports by a human expert. Our main goal was to capture information contained in the reports rather than to reflect all complexities of the domain.

The organization of the paper is as follows. First, the general system architecture is presented, then, a partial description of adopted ontology is provided and a sample extraction rule is given. Next, merging procedures for specific blocks are outlined and their partial evaluation is presented. The final section contains conclusions.

## 2 System Architecture

As stated in [6], the system is a pipeline of simpler modules, dedicated to solving specialized tasks: pre-processing, getting partial information by SProUT IE rules, cleaning and merging the results, see Fig. 1. Currently, the merging procedure has been split into two phases: the annotation of report's main blocks (breast's

composition, findings, general diagnosis and recommendations) and then segmenting data in these blocks (e.g., type of dominant and remnant tissue, their localization or concentration).



**Fig. 1.** System architecture

The first two stages (pre-processing and IE) remain essentially the same as reported before ([6], [5]). The clean-up module has been simplified as variable coreference is done automatically with the new version of SProUT. Thus, the module is responsible only for deleting morphological information from output structures, removing duplicate analyses and performing pseudo-unification of localizations (see [6] for details). As mentioned above, merging has been currently separated into two phases: first, general blocks are recognized, and then data within these blocks are segmented so that they fully correspond to structures which will be stored in the final database. The latter phase was mostly motivated by data in the breast’s composition block. Information about the type of dominant tissue is often interleaved with localization of the remnant tissue and a sequential annotation is impossible. Therefore, an additional procedure is introduced which linearizes and groups the results, see sec. 5.2.

### 3 Ontology

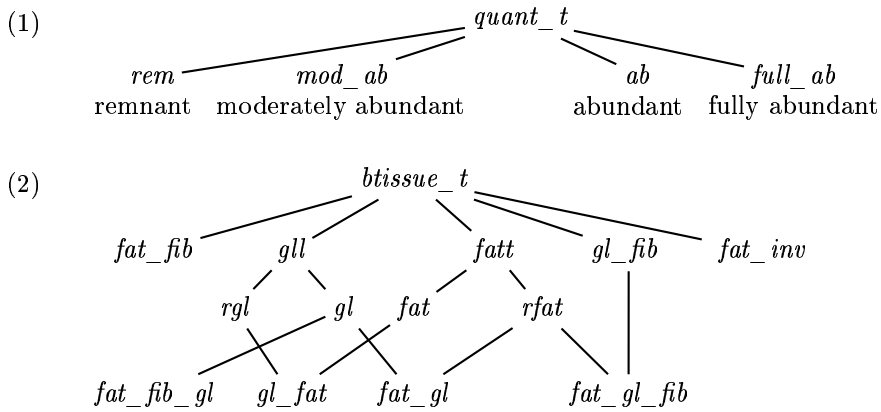
On the basis of the Polish mammographic ontology developed by T. Podsiadly-Marczykowska, we prepared a simplified model adjusted to our needs, and represented as attribute-value pairs (AVMs). The attributes representing general information about examination, findings and localization have been described in [6]. In Table 1, we present attributes used in the representation of the breast’s composition.

Values of most attributes are the direct subtypes of their most general type and form a shallow type hierarchy; e.g., the value of GLAND|QUANT is of the *quant\_t* type, with subtypes given in (1). For some values, however, it is convenient to specify more complex relations in order to take advantage of strong typing and unification in the current version of SProUT (3.7.2). The hierarchy of tissue types uses this kind of complex constraints, see (2). Names of tissue types, with the exception of the glandular and fat tissue, consist of two elements: the dominant tissue type (the first element of the name) and the remnant tissue (the

breasts' description		
BRSURG	SURGERY	undergone surgeries
BRSURG	REASON	the reason of the surgery
BTISSUE		main type of the breast tissue
FIBRE		fibrosis
CHAR		features of breast tissue, e.g., displastic or difficult for evaluation
description of glandular tissue		
GLAND	QUANT	quantity of glandular tissue
GLAND	REGULAR	regularity of glandular tissue
GLAND	DENSITY	density of glandular tissue
GLAND	MACULATION	maculation of glandular tissue
comparison of glandular tissue		
CMP	REG	comparison of regularity
CMP	DENSITY	comparison of density
CMP	QUANT	comparison of quantity
general attributes		
LOC		tissue localization
DIAGNOSIS_RTG		radiological diagnosis of breast's composition
RECOMMENDATION		recommendation

**Table 1.** Attributes for the representation of breast's composition and their meaning

remaining part). For example, if the remnant tissue is glandular-fibrosis (*gl\_fib*) and the dominant type is the fat tissue, the resulting tissue is *fat\_gl\_fib*.



In (2), the hierarchy of tissue types is presented. The most general type is *btissue\_t*. The only types which cannot be the most specific are the direct subtypes of the glandular (*gll*) and fat (*fatt*) tissue, i.e., *rgl*, *gl*, *rfat*, *fat*. The types *rgl* and *rfat* denote dominant tissue types. The other two types represent the remnant tissue. Tissue types can appear directly in the text, e.g., *o tkance tłuszczowo-gruczołowej* 'with fat-glandular tissue' or can be combined from two

separate types, e.g., *resztkowa tkanka gruczołowa z przewagą tłuszczowej* ‘remnant glandular tissue with dominating fat tissue’. In both cases, the resulting type is *gl\_fat*. The hierarchy in (2) does not reflect all possible combinations of tissue types but only those which follow from the domain specification.

## 4 Breast Tissue Extraction Rules

The rule in (3) recognizes different phrases describing the main tissue type if the dominant tissue type is stated in the text and there is no information about the remnant type. We decided that the default remnant tissue type is glandular (represented by *gl*). This decision follows from the general medical knowledge and the analysis of mammogram reports.

```
(3)  t_majority:>
      (@seek(loc) & [LOC #loc])?
      (morph & [POS prep, SURFACE ‘o’, INFL infl_prep &
              [CASE_PREP #c1]]
      @seek(tiss)& [C #c1])?
      (morph & [POS prep, SURFACE ‘z’, INFL infl_prep &
              [CASE_PREP ins]])?
      (morph & [STEM ‘przewaga’] | morph & [STEM ‘przeważać’])
      (@seek(tiss)& [C gen])?
      (gazetteer & [GTYPE gaz_med_utkanie, G_CONCEPT #typ])
      -> btiss_str & [BTISSUE gl & #typ, LOC #loc].
```

The `t_majority` rule<sup>1</sup> captures, among others, the following phrases: *sutki o utkaniu z przewagą tłuszczowego* or *sutki z przewagą tkanki tłuszczowej*, both meaning ‘breasts with the dominant fat tissue’. The type of the dominant tissue is recognized by a gazetteer entry, indicating a tissue type ([GTYPE *gaz\_med\_utkanie*]) with a specific G\_CONCEPT value. The words *tłuszczowego* or *tłuszczowej* ‘fat’ are represented in the gazetteer by the concept *fatt* (the G\_CONCEPT value). The variable `#typ` transports information about the tissue type to the resulting structure, where it is unified with the *gl* type (the default value of the remnant tissue). As the result of unification of these two values, we obtain the *fat\_gl* type (see (2)).

## 5 Automatic Annotation

The IE results are stored in a text file as a sequence of attributes and their values. In order to separate the main blocks (findings, breast’s composition and diagnosis) and their components (e.g., if there are several findings, their scope has to be identified), we automatically insert tags indicating beginning/end of each element. The next three subsections present the relevant procedures.

<sup>1</sup> Details of the SProUT rules’ syntax can be found, e.g., in [7].

## 5.1 Findings

In order to separate findings, we insert tags identifying the beginning (zp) and end (zk) of each finding. The algorithm presented in [6] remains unchanged (except for updating attributes' names) and is briefly outlined below.

According to their content, attributes are divided into 4 classes concerning: 1) mammographic findings, 2) breast's composition, 3) an overall report evaluation, and 4) attributes used in descriptions of both findings and breast's composition. The annotation procedure is build around one of the two attributes which unequivocally identify the finding: ANAT\_CHANGE (anatomic change) or INTERPRETATION. Starting with the first such attribute, we are trying to cover the maximal part of the report until attributes from a different block appear or attributes unique for the finding block are repeated. The final zp and zk boundaries can be corrected if any localization attributes remain unattached to any other block. Sample annotation results are given in (4).

```
(4)  bp
      EXAM_ID:32742|PATIENT_ID:41590
      up
      LOC|BODY_PART:breast||LOC|L_R:left-right
      QUANT:rem
      BTISSUE:g11
      LOC|LOC_CONV:uoq
      BTISSUE:fat_g1
      uk
      zp
      LOC|BODY_PART:breast||LOC|L_R:left-right
      GRAM_MULT:number||MULT:single
      ANAT_CHANGE:macro||GRAM_MULT:plural
      DIAGNOSIS_RTG:benign
      zk
      DIAGNOSIS_RTG:no_susp||LOC_D|BODY_PART:armpit||LOC_D|L_R:left-right
      rp
      PREV_EXAM|MONTH:1_||SATURATION_CHANGE:_none|SIZE_CHANGE:_none
      bk =====
```

## 5.2 Breast's Composition

The next annotation step is to divide the breast's composition block (delimited by up/uk tags) into logical subblocks referring to the AVM attributes in (5).

```
(5)  [
      BRSURG brsurg_avm
      ...
      RECOMMENDATION recommendation_t
      MAIN_TISSUE [ BTISSUE btissue_t
                   LOC loc ]
      GLAND_TISSUE ( [ BTISSUE btissue_t
                      LOC loc
                      GLAND_ATTR gl_attr_avm ], ... )
      COMPARISON ( [ CMP_ATTR cmp_attr_avm
                    LOC loc ], ... )
    ]
```

The breast's composition block is further annotated as follows. Lines containing general attributes: SURGERY, REASON, CHAR, DIAGNOSIS\_RTG, RECOMMENDATION or EXAM\_ID are marked as `info`. Lines containing the LOC attributes are marked as: 1) `log` if there is only general localization information, e.g., breast and lateralization; 2) `lsz` if there are only attributes denoting an anatomic part or a conventional localization, 3) `loc` if both `loc` and `lsz` attributes are present.

A line containing the attribute BTISSUE of type *gll* or *gl\_fib* is marked `tsz`; for other types the marker is `tog`. If a line contains a GLAND attribute, we mark it as `gland` and `CMP_*` attributes are tagged with `cmp`.

The breast's composition block is processed from the end, lines marked as `info` are left unchanged and the subblock's closing tag, `utk`, is inserted. Then, we process the block according to one of the following rules: a) lines marked with `log` or `loc` remain unchanged. When it is `lsz`, we look upward for the line marked `log` and copy it after `lsz` (a kind of localization unification). Then we recognize a block of consistent information; for simplicity, we assume that it corresponds to lines marked `tog`, or `tsz`, or `gland` with `tsz`, or `cmp` with `tsz`; b) a line marked `tog` is left unchanged; then we look upward for the nearest `log` and add this localization to the `tog` tissue (unification); c) a block of consistent information is recognized and then the localization with unification are added to it.

Next the subblock's opening tag (`utp`) is inserted and we repeat the operation of determining a subblock.

The breast's composition block in (4) refers to the text: *Sutki o resztkowym utkaniu gruczołowym w kwadrantach górno-zewnętrznych. Przewaga tkanki tłuszczowej.* 'Breasts with the remnant glandular tissue in upper-outer quadrants. The dominant fat tissue.' Our algorithm transforms this block into (6):

```
(6)      LOC|BODY_PART:breast||LOC|L_R:left-right
         utp
         GLAND|QUANT:rem
         BTISSUE:gll
         LOC|BODY_PART:breast||LOC|L_R:left-right
         LOC|LOC_CONV:uoq
         utk
         utp
         LOC|BODY_PART:breast||LOC|L_R:left-right
         BTISSUE:fat_gl
         utk
```

### 5.3 Overall Diagnosis and Reliability

We have used three attributes to provide a compact summary of the report: `REPORT_CLASS` (for the overall diagnosis), `MMG_REL` (to indicate how reliable the image is) and `REPORT_WITH_FINDINGS` (a binary distinction specifying if any findings have been detected). The value of `REPORT_CLASS` is inferred from component diagnoses (if any) and recommended examinations: the most severe diagnosis is taken for the overall diagnosis unless an oncological consultation

is required, which yields the *diag\_mal* (malicious) value, or a biopsy is recommended, which results in the *diag\_susp* (suspicious) value.

The reliability of the diagnosis depends on the type of the breast's composition. The values of the MMG\_REL attribute are assigned as follows: if the breast tissue is very dense or dysplastic, or it is explicitly stated that the image is difficult for evaluation, MMG\_REL is *unreliable*. If the fat tissue is dominant, the report is *reliable*; in all other cases, MMG\_REL takes the *avg\_reliable* value.

## 6 Evaluation

The evaluation was done on 705 new reports. The results of the automatic annotation were checked manually. We tagged all places where any feature or block marking was inserted incorrectly, was not inserted or was inserted in a wrong place. Afterwards we counted all correct, misplaced and incorrect occurrences of all attributes. A selected part of the results is presented in Fig. 2. The main issue was the proper recognition of the end of the composition block. Several times the last localisation attribute was incorrectly attached to the composition block instead of a finding block. This was the primary reason for correcting the boundaries of these blocks. Another important issue is lack of a clear distinction between two main concepts specifying finding blocks: an anatomical change, which can occur at most once in a finding block, and interpretation, which can appear several times in the block. Some errors were also caused by incomplete grammar coverage for negation and comparison phenomena. For example, a phrase *a density seen in the previous examination now is not visible* was incorrectly recognized as identifying a finding description. On the other hand, a finding described by the phrase *the biggest of them (was located in...)* was not recognized.

## 7 Conclusions

We have presented a system for automatic processing of Polish mammogram reports. In particular, we focused on obtaining information about radiological findings, breast's composition, an overall diagnosis and its reliability. Since the extracted data is normalized, it allows for an easy access to similar cases and can support physicians in diagnosing or comparing the data. The evaluation results are encouraging and indicate that the presented method is quite successful: for the most complex task, i.e., grouping several attributes into blocks, we obtained about 82% accuracy, and even 10% higher for single attributes. We believe that the main problems which caused the accuracy decrease can be relatively easy to eliminate in case of a practical application of the system.

## References

1. Busemann S. and Krieger H.-U. Resources and Techniques for Multilingual Information Extraction. In: *Proceedings of LREC 2004, Lisbon, Portugal, 2004*, pp. 1923–1926.

patient records	705	%
FINDINGS	338	100
unrecognized findings	17	5.0
incorrectly recognized findings	34	10.1
correctly recognized positions of block beginnings	275	81.4
incorrectly recognized positions of block beginnings	46	13.6
incorrectly recognized positions of block endings	27	8.0
BREAST COMPOSITION SUBBLOCK	968	100.0
incorrectly recognized subblocks	9	1.0
unrecognized subblocks	3	0.3
incorrectly recognized positions of subblock endings	24	2.5
ANAT_CHANGE	276	100.0
correctly recognized	269	97.5
incorrectly recognized	22	0.8
INTERPRETATION	174	100.00
correctly recognized	163	93.7
incorrectly recognized	3	1.7

**Fig. 2.** Evaluation of automatically identifying blocks' boundaries/attributes

2. Drożdżyński W., Krieger H.-U., Piskorski J., Schäfer U., and Xu F. Shallow Processing with Unification and Typed Feature Structures — Foundations and Applications. In: *German AI Journal KI-Zeitschrift*, 01/04. Gesellschaft für Informatik e.V, 2004.
3. Jain N. L. and Friedman C. Identification of Findings Suspicious for Breast Cancer Based on Natural Language Processing of Mammogram Reports. In: *Proceedings of the American Medical Informatics Association Annual Fall Symposium*, 1997, pp. 829-833.
4. Hahn U., Romacker M., and Schultz S. MEDSYNDIKATE — a natural language system for the extraction of medical information from findings reports. In: *International Journal of Medical Informatics*, 2002, pp. 63-74.
5. Kupść A., Marciniak M., Mykowiecka A., Piskorski J., and Podsiadły-Marczykowska T. Information Extraction from Mammogram Reports. In: *KONVENS 2004*, Vienna, Austria, 2004, pp. 113–116.
6. Marciniak M., Mykowiecka A., Kupść A., and Piskorski J. Intelligent Content Extraction from Polish Medical Reports. In: *Proceedings of ICMIT 2004*, Warsaw, Poland, 2004.
7. Piskorski J., Homola P., Marciniak M., Mykowiecka A., Przepiórkowski A., and Woliński M. Information Extraction for Polish using the SProUT Platform. In: *Proceedings of ISMIS 2004, Zakopane*, 2004, pp. 225–236.
8. Ruch P., Baud R., and Geissbruhler A. Evaluating and reducing the effect of data corruption when applying bag of words approaches to medical records. In: *International Journal of Medical Informatics*, 2002.