Intelligent Content Extraction from Polish Medical Reports

Małgorzata Marciniak[1], Agnieszka Mykowiecka[1], Anna Kupść[1] and Jakub Piskorski[2]

[1] Institute of Computer Science, Polish Academy of Sciences, Ordona 21, 01-237 Warsaw, Poland
[2] DFKI GmbH, Stuhlsatzenhauseweg 3, D-66123 Saarbuecken, Germany

Abstract. The paper presents a method for intelligent automatic processing of medical reports. First, we extract single pieces of information using SProUT (a general-purpose IE platform), and then, externally merge the results in order to obtain a detailed formalized description of the reports.

1 Introduction

The paper presents a method for intelligent automatic processing and understanding of Polish medical texts. In particular, we aim at standardizing and formally representing data from mammogram reports so that they can be stored in a database which can support physicians in decision making and diagnosing. (A similar task for English mammogram reports was undertaken by [2] and [3].) Our processing method consists of two phases: simple template extraction and merging these templates into more complex structures.

We obtain single pieces of information using SProUT [1] which has been adapted to processing of Polish [5]. The extracted information is represented as attribute-value pairs, according to the domain model specified in [4]. The model comprises a simplified ontology which represents main mammographic concepts. The extracted data is normalized and linked to the ontology elements, which will enable concept-driven querying of the final database.

In the second phase, we merge the results in order to obtain more complex structures defined in the model. This phase consists of cleaning up the results, performing pseudo-unification and grouping the attributes so that they describe appropriate objects. The latter operation is crucial, as a single report often contains several mammographic findings or other objects of the same type. In order to separate their descriptions, we have implemented several heuristics and the separators are inserted automatically in the processed reports. The initial evaluation of the algorithm is presented in sec. 4.

2 SProUT Extraction Rules

SProUT is a multilingual NLP platform equipped with a set of reusable Unicode-capable online processing components for various linguistic operations which can

be flexibly combined into a pipeline that produces several streams of linguistically annotated structures. These structures serve as an input for the shallow grammar interpreter, applied at the next stage. The grammar formalism in SProUT combines efficient finite-state techniques and expressive unification-based formalisms. To be more precise, a grammar in SProUT consists of pattern/action rules, where the LHS of a rule is a regular expression over TFSs with functional operators and coreferences, representing the recognition pattern, and the RHS of a rule is a TFS specification of the output structure. Coreferences express structural identity, create dynamic value assignments, and serve as means of information transport. Functional operators are utilized for forming the output of the rules and for introducing complex constraints.

Currently, our SProUT grammar consists of 211 extraction rules. Each rule detects words or phrases which describe findings, their features, interpretation(s) or recommended further examinations. For example, the rule below recognizes different preposition phrases identifying calcifications accompanying the finding. Micro- and macrocalcifications and their quantity are recognized by the `multiplicity` rule which is called (via the `@seek` operator) by the `with_calc` rule. The quantity of calcifications is either expressed directly, e.g., *bardzo liczne* 'numerous', *nieliczne* 'a few', etc., or by the grammatical number (singular/plural) of a word corresponding to the calcification. The detected calcification type and its quantity are concatenated by the functional operator `ConcWithBlanks` and stored in the resulting output structure.

```
with_calc:>
      ((morph & [POS prep, STEM 'z', INFL [CASE_PREP ins]] |
        morph & [POS prep, STEM 'ze', INFL [CASE_PREP ins]])
       (morph & [STEM 'obecność'])?
       (@seek(multiplicity) & [MULT #mu])?
       (gazetteer & [GTYPE gaz_med_zmiana, CONCEPT mikrozwapnienie & #zm, G_NUMBER #li]|
        gazetteer & [GTYPE gaz_med_zmiana, CONCEPT makrozwapnienie & #zm, G_NUMBER #li]|
        morph & [STEM 'zwapnienie' & #zm, INFL[ NUMBER_NOUN #li]]) )
->[WITH_CALCIF #zm, C_MULT #mul],  where #mul = ConcWithBlanks(#mu,#li) .
```

3 Post-Processing and Automatic Annotation

Every phrase recognized by SProUT is stored as a disjunct, with alternative analyses if more than one grammar rule applied to the phrase. In postprocessing, the resulting output structure for each rule is obtained and cleaned up.

First, all variable coreferences are resolved. In many cases, an attribute's value is given indirectly, by a reference to a different value. Therefore, we substitute all references with the corresponding values. Then, duplicate analyses are removed. Grammar rules often output morphological information (case, number and gender) in order to ensure agreement between elements in the identified phrases. Maintaining this information in the output structures usually produces alternative analyses. As morphology is irrelevant for our final results, it is deleted from the output structures and duplicate analyses are removed. Finally, in order to simplify annotation heuristics, we perform a pseudo-unification on some attribute-value pairs. Although we attempt at collecting all information about localization in a single structure, it is not always possible. Usually different pieces

of information about localization are recognized one after another but, due to numerous paraphrases, they are separated by miscellaneous text and extraction rules cannot group them in a single structure. If the two adjacent pieces of information are complementary, e.g., the first recognized localization concerns a body part and lateralization, whereas the second lacks this information but specifies conventional localization, the two pieces are merged into a single structure.

The postprocessed report is a text document containing a sequence of attribute-value pairs. If the output structure consists of several attribute-value pairs, as in (1b), they are treated as one result in the document. Each result is stored in one line, with attributes separated by '||', i.e., (1a)–(1b) correspond to (2).

(1) a. $\left[\text{EXAM\_ID } 237\right]$ b. $\begin{bmatrix}\text{RECOMMENDATION USG}\\ \text{TIME za rok}\end{bmatrix}$

(2)
```
EXAM_ID:237
RECOMMENDATION:USG||TIME:za rok
```

The most complex structures in the adopted mammography model are descriptions of breasts' composition and findings. In order to delimit their descriptions in processed reports, we introduce several tags and use a few heuristics to insert them at the beginning and at the end of each block. The two main blocks are indicated by `up` (`uk`), start (end) of the breast's composition description, and `zp` (`zk`) tags, start (end) of a finding description.

The annotation of each report is built around the attributes representative for each block, i.e., ANAT_CHANGE, INTERPRETATION (for findings), and BTISSUE (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.

4 Evaluation

The evaluation consisted in checking how many findings were identified by our algorithm and how many of their attributes were adequately recognized. We used for evaluation 448 reports from two different health care providers, which constituted (a random) 33% of all reports used for the system development. For evaluation, we selected only reports describing some irregularities. We processed them automatically and then manually examined the obtained results. We marked all incorrectly inserted tags or attributes as '−', whereas all missing tags and attributes were preceded by '+'. Then, we compared the original results with their manual corrections, counting the number of incorrectly recognized attributes and findings' boundaries. The obtained results are presented in Fig. 1.

The main reasons of detected errors are: 1) coordination — some elements 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 causes opposite interpretations; 3) paraphrases — different ways of expressing the same concept.

	nb	%
patient records	448	
FINDINGS	474	100
findings with correctly recognized beginnings	416	87.76
unrecognized findings	13	2.74
incorrectly recognized findings	17	3.59
findings with incorrectly recognized beginnings	45	9.50
sample attributes:		
SATURATION	185	100
correctly recognized	182	98.38
WITH_CALCIF	40	100
correctly recognized	35	87.50

Fig. 1. Evaluation of Findings' Descriptions

5 Conclusions

Although the approach described in the paper turned out to be quite effective, there is still room for improvement. Our future work will focus on: a) writing more complex grammar rules to cover problematic cases (e.g., different types of relations in comparisons or coordination); b) deploying more fine-grained grouping heuristics; c) applying machine learning techniques to obtain paraphrases and extraction patterns; d) supporting processing with an inference mechanism. The latter enhancement 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.

References

1. 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.
2. N. L. Jain and Carol Friedman, 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.
3. U. Hahn, M. Romacker and S. Schultz, MEDSYNDIKATE — a natural language system for the extraction of medical information from findings reports. In: *International Journal of Medical Informatics*, 2002, pp. 63-74.
4. A. Kupść, M. Marciniak, A. Mykowiecka, J. Piskorski, and T. Podsiadły-Marczykowska. 2004. Information Extraction from Mammographic Reports. In: *KONVENS 2004*, Vienna, Austria.
5. J. Piskorski, P. Homola, M. Marciniak, A. Mykowiecka, A. Przepiórkowski, and M. Woliński. 2004. Information Extraction for Polish using the SProUT Platform. *Proceedings of ISMIS 2004, Zakopane.*

This article was processed using the LaTeX macro package with LLNCS style