Making Shallow Look Deeper: Anaphora and Comparisons in Medical Information Extraction

Agnieszka Mykowiecka*, Małgorzata Marciniak*, Anna Kupść†*

*Institute of Computer Science, Polish Academy of Sciences
Ordona 21, 01-237 Warsaw, Poland
{agn,mm}@ipipan.waw.pl
†Langue et Dialogue, Loria
BP 239 – Campus Scientifique, 54506 Vandoeuvre-lès-Nancy, France
kupsc@loria.fr

Abstract

The paper focuses on resolving natural language issues which have been affecting performance of our system processing Polish medical data. In particular, we address phenomena such as ellipsis, anaphora, comparisons, coordination and negation occurring in mammogram reports. We propose practical data-driven solutions which allow us to improve the system's performance.

1. Introduction

The aim of the paper is to address a few pertinent issues which we stumbled upon while developing the system described in (Marciniak et al., 2004). The main goal of the system is to structure data encoded in medical reports and then store it in a database[1] which will serve as a starting point for either quantitative or qualitative study.

The system processes medical texts written in Polish[2] and is based on Information Extraction (IE) techniques (shallow parsing and keyword recognition), which enables more efficient data processing than a deep text analysis. Although shallow methods are in principle sufficient for this task, natural language phenomena such as ellipsis, comparisons, coordination, anaphora or negation affect quality of the system's results. In this paper we present a few practical solutions, tailored for this application.

The organization of the paper is as follows: first, we briefly present the system and its components; then, linguistic phenomena related to the analysis of three specific types of information recognized by the system (localization of mammographic findings, anatomical changes and comparison with previous examinations) are discussed, and the solutions adopted in the system are presented. The paper closes with a demonstration of the effect the improvements have on the system's performance.

2. The IE System

The system processes Polish mammogram reports which are brief and compact texts accompanying radiological images. The core of the system is SProUT, a general-purpose IE platform (Becker et al., 2002) adapted for Polish (Piskorski et al., 2004). Data structures in SProUT are uniformly represented as attribute-value pairs (AVM, Attribute-Value Matrix), which allows for a powerful blend of finite-state techniques (regular expressions) and a unification-based formalism expressiveness.

Data processing has been split in the system into several modules: pre-processing (correcting spelling and punctuation errors in the texts), IE using SProUT, post-processing (removing duplicate analyses, deleting irrelevant information, etc.), and extracted data aggregation according to a domain conceptualisation model proposed by (Kupść et al., 2004).

The domain model is AVM-encoded and represents only mammographic concepts relevant for the processing task: breast's composition, types of tissue, finding, diagnosis, recommended examinations, etc. Most concepts are organized into a relatively flat hierarchy but some of them (e.g., tissue types) form a complex multi-inheritance structure. The subsumption relation is heavily used by the unification mechanism in SProUT rules.

The merging procedure consists in: 1) automatic annotation of the report's main blocks (breast's composition, findings, general diagnosis and further examinations) and 2) segmenting data in these blocks (especially the breast's composition block requires further annotation of the type of dominant and remnant tissue, their localization and concentration). At the final stage, a compact summary of the report is also provided. It is represented by three separate attributes, corresponding to an overall diagnosis, its reliability and an indication if any findings have been detected. The values of the summary attributes are inferred from the extracted data rather than obtained directly from the report. A sample report with the corresponding processing results is presented in Fig. 1. The tags indicate: `bp` and `bk`, respectively, the beginning and end of the report; `up` and `uk` delimit the breast's composition block; tissue type and its localization are additionally indicated in this block by the `utp` and `utk` tags; finding descriptions are annotated with `zp` and `zk`. As mentioned above, the last three attributes of the report (MMG_REL, REPORT_CLASS and REPORT_WITH_FINDINGS) are used for classification purposes.

The system has been evaluated in (Mykowiecka et al., 2005) and the results are quite encouraging: about 82%

[1] The database storage phase is currently under development.

[2] Although we are not aware of other research in this area concerning Polish data, we can point to various efforts undertaken for English. English mammogram reports were dealt with by (Jain and Friedman, 1997) and (Hahn et al., 2002) while chest radiology reports were processed by (Hripcsak et al., 2002), (Taira et al., 2001) and (Hripcsak et al., 1995).

775 *Sutki o utkaniu z przewagą tłuszczowego. W sutku prawym przybrodawkowo widoczny guzek o śr. 10mm z makrozwapnieniami w jego obrębie odpowiadający f-a degenerativa (zmiana łagodna).*

Breasts with the dominant fat tissue. In the right breast in subareolal, there is a tumor of 10mm diameter with macrocalcifications corresponding to f-a degenerativa (benign finding).

```
bp
   EXAM_ID:775
up
   LOC|BODY_PART:breast||LOC|L_R:left-right
utp
   LOC|BODY_PART:breast||LOC|L_R:left-right
   BTISSUE:fat_gl
utk
uk
zp
   LOC|BODY_PART:breast||LOC|L_R:right
   ANAT_CHANGE:mass||GRAM_MULT:singular
   DIM:mm||NUM1:10||NUM2:10
   C_GRAM_MULT:plural||WITH_CALC:macro
   INTERPRETATION:f-a_deg
   DIAGNOSIS_RTG:benign
zk
   MMG_REL:reliable
   REPORT_CLASS:diag_benign
   REPORT_WITH_FINDINGS:yes
bk
```

Figure 1: A sample mammogram report

accuracy for annotation of complex blocks (findings) and about 10% more for simple templates. However, as admitted in the evaluation, the system's performance is not optimal due to problems such as negation, comparisons or coordination. At present, we are going to pay closer attention to these issues and propose amendments to improve the results.

3. Identified Problems

Although the system was able to distinguish mammograms of healthy and ill patients quite well, not all findings have been properly recognized. As this information seems crucial for a statistical data analysis, the reasons of the most typical errors should be examined. Some problems were caused by lack of entries in the domain lexicon or an insufficient coverage of grammar rules and are quite easy to eliminate. However, the issues listed below are much harder to resolve:

- anaphoric expressions in localizations and anatomic changes;
- a representation of coordination for localizations and anatomic changes;
- relating the current report to the previous one.

In the paper, we propose pragmatic, data- and domain-driven solutions to these issues. The subsequent sections will deal with, respectively, descriptions of localizations, anatomic changes and data referring to previous examinations.

4. Localization

Localization information is important and very frequent in mammogram reports. In a sample of 705 reports, there were 2196 localization statements; 25 of them were not recognized by the system while 4 were incorrectly recognized. The overgenerated localizations were all caused by an improper analysis of information concerning previous examinations and will be addressed in sec. 6. Presently, we will focus on remedying undergeneration of localization phrases.

4.1. Localization Anaphora

By a 'localization anaphora' we understand phrases where localization is specified by a reference to a previously used expression. For example, if there are two findings identified in a breast, lateralization information (i.e., left or right) is usually omitted, e.g. *w kwadrancie górno-zewnętrznym tego sutka* 'in the upper-outer quadrant (uoq) of this breast', *w tym sutku w KGZ* 'in this breast in uoq'. In these phrases, the demonstrative pronouns *tym, tego* ('this$_{inst/gen}$') should be interpreted as 'in the same breast which has been described previously'. Similarly, the expression *sąsiadujący* 'neighbouring' can be used.

We have introduced new subtypes for values of all attributes in the LOC structure (i.e., BODY_PART, L_R, LOC_CONV etc.). When the above mentioned phrases are found, the path LOC|L_R is assigned the new *loc_l_r_last* type, (1), which is replaced with the appropriate localization information in the merging phase.

(1)
```
LOC|BODY_PART:breast||LOC|LOC_CONV:liq
  ||LOC|L_R:loc_l_r_last
```

4.2. Coordination of Localizations

In certain reports, some information (e.g., a finding) concerns several localizations, as in (2). In previous experiments such information was improperly associated with only one (the last) localization. Information about other localizations was lost.

(2) *W kwadrantach górno-zewnętrznych obu sutków oraz w okolicy zabrodawkowej sutka prawego pojedyncze wyraźnie okonturowane zagęszczenia o śr do 5 mm. Zmiany łagodne (wewnątrzsutkowe węzły chłonne? torbielki?)*

In the upper-outer quadrants and in subareolal of the left breast, there are single well circumscribed densities of a diameter up to 5 mm. Benign changes (intramammary lymph nodes? cysts?)

In order to deal more adequately with such cases, we introduce a new grammar rule which can cover several localizations and which assigns values to the LOC attribute as well as to the newly added LOC2 and LOC3 attributes. In the merging phase, lines with LOC2 and LOC3 values are separated and copies of the block where they occur (without LOC2 and LOC3 attributes) are created, compare (3) and (4).

(3)
```
bp
EXAM_ID:26699||PATIENT_ID:38696
zp
 LOC|BODY_PART:breast||LOC|LOC_CONV:uoq||LOC
   |L_R:left-right
 LOC2|BODY_PART:breast||LOC|LOC_CONV1:subareolar
   ||LOC|L_R:right
 MULT:single
 CONTOUR:circumscribed
 ANAT_CHANGE:density||GRAM_MULT:plural
```

```
 DIM:mm||NUM1:5
 DIAGNOSIS_RTG:benign
 INTERPRETATION:intr_lymph_node
 INTERPRETATION:cyst
zk
```

(4)
```
bp
 EXAM_ID:26699||PATIENT_ID:38696
zp
 LOC|BODY_PART:breast||LOC|LOC_CONV:uoq||LOC
   |L_R:left-right
 MULT:single
 CONTOUR:circumscribed
 ANAT_CHANGE:density||GRAM_MULT:plural
 DIM:mm||NUM1:5
 DIAGNOSIS_RTG:benign
 INTERPRETATION:intr_lymph_node
 INTERPRETATION:cyst
zk
zp
 LOC|BODY_PART:breast||LOC|LOC_CONV1:subareolar
   ||LOC|L_R:right
 MULT:single
 CONTOUR:circumscribed
 ANAT_CHANGE:density||GRAM_MULT:plural
 DIM:mm||NUM1:5
 DIAGNOSIS_RTG:benign
 INTERPRETATION:intr_lymph_node
 INTERPRETATION:cyst
zk
```

The same grammar rule is used when one diagnosis (DIAGNOSIS_RTG) is associated with more than one localization, (5). In the postprocessing phase, the information as in (6) is split into two or three lines containing a separate diagnosis for each localization.

(5) *Doły pachowe, skóra i brodawki sutkowe wolne.*

Armpits, skin and nipples are non-malignant.

(6)
```
DIAGNOSIS_RTG:no_susp||
 LOC2_D|BODY_PART:breast||LOC2_D|LOC_A:skin
       ||LOC2_D|L_R:left-right||
 LOC3_D|BODY_PART:breast||LOC3_D|LOC_A:nipple
       ||LOC3_D|L_R:left-right||
 LOC_D|BODY_PART:armpit||LOC_D|L_R:left-right
```

5. Anatomic Changes

Another important information which was not satisfactorily recognized was the occurrence of anatomic changes.

Some findings are described in mammogram reports by elliptic expressions, as in the following phrases: *podobna zmiana* 'a similar finding', *druga* 'the second (one)', or *zmiana o tej samej wielkości i charakterze* 'a finding of the same size and type (as the previous one)'. In such cases, the system should refer to the preceding finding in order to obtain the missing information. Let us consider a few examples.

(7) *Sutek prawy – na pograniczu kwadrantów dolnych widoczne wyraźnie ograniczone zacienienie o śr ok 10 mm. Zmiana radiologicznie łagodna (torbielka?). Sutek lewy – zmiana podobnej wielkości i charakteru w KDZ.*

In the right breast, at the border line of lower quadrants, there is a well circumscribed darkness of about 10 mm diameter. A benign finding (cyst?). The left breast – a finding of the similar size and type in the lower outer quadrant.

The system should detect two finding blocks in the report in (7). The first finding is described precisely. For the second one, however, only its localization is specified (the upper outer quadrant of the left breast) while other properties of the finding (type and size) are stated by a reference to the previous one. Hence, the system should obtain the relevant data from the previous finding, i.e., ANAT_CHANGE darkness, CONTOUR circumscribed, DIAGNOSIS_RTG benign and INTERPRETATION cyst, as well as SIZE. A similar situation occurs in (8) where the second finding is mentioned by the elliptic phrase *drugie* 'the second (one)'. Localization and size of the finding are given explicitly but other information has to be inferred from the previous finding's description.

(8) *W sutku prawym w KGW 5mm dobrze ograniczone zagęszczenie (łagodne), drugie w KGZ o śr. 10mm również o podobnym charakterze.*

In the right breast, in the upper inner quadrant, there is a well circumscribed density (benign), the second one in the upper outer quadrant, of 10mm diameter size and a similar type.

In order to solve such problems, we introduced three attributes (CHANGE_REF, and two boolean attributes: TYPE_REF, SIZE_REF) which indicate referential values of the selected attributes. The attribute CHANGE_REF has two possible values *ref_cmp* and *ref_max*. The first one indicates that a new finding has been identified and its description shares some information with the previous finding block. The second value does not indicate a new change but refers to the biggest finding in the previous block (see (10) below). The (*yes* value of the) TYPE_REF shows that the finding description is the same as the description of a finding in the previous block. Similarly, SIZE_REF specifies that the size of the finding is identical to the value of SIZE or SIZE_TEXT (whichever is present) in the previous finding block.

To identify phrases indicating similar findings we introduced several specific grammar rules based on the occurrence of certain words. The system output obtained for example (8) is given in (9).

(9)
```
zp
   LOC|BODY_PART:breast||LOC|LOC_CONV:uiq
                          ||LOC|L_R:right
   DIM:mm||NUM1:5
   ANAT_CHANGE:density
        ||CONTOUR:circumscribed
        ||GRAM_MULT:singular
   DIAGNOSIS_RTG:benign
zk
zp
   CHANGE_REF:ref_cmp
   LOC|LOC_CONV:uoq
   DIM:mm||NUM1:10||NUM2:10
   TYPE_REF:yes
zk
```

If there are several anatomic changes and more detailed information is provided only for one of them, as in (10), the remaining properties are shared with other findings. The referential expressions are usually preceded by the word *największy* 'the biggest (one)' and are detected by grammar rules which add to the output structure the attribute CHANGE_REF with the *ref_max* value.

(10) *W obu sutkach różnej wielkości gładkokonturowane zagęszczenia o charakterze zmian łagodnych radiologicznie (torbielki? f-a?). Największa*

zmiana o śr ok 12 mm widoczna na pograniczu kwadrantów górnych sutka lewego.

In both breasts, there are smoothly circumscribed densities of different sizes, benign character (cysts? f-a?). The biggest finding of 12 mm diameter is visible at the border line of upper quadrants in the left breast.

6. Findings' Progress: Comparison with Previous Examinations

Radiologists often refer to the previous examination in order to verify findings' progress or identify new changes. This part of the report turned out to be quite problematic for the text analysis.

The major issue was the incorrect identification of findings which were no longer present. For example, in (11) negation has not been treated properly and the system recognized *zagęszczenie* 'density' as still present (identified fragments are indicated in square brackets).

(11) *Opisywane [w badaniu poprzednim z dnia 22. 12.98r] [zagęszczenie] [w sutku prawym] obecnie nie jest widoczne.*

[The density] [in the right breast], described [in the report from 22.12.98], is no longer visible.

A similar construction is often used to specify that the finding's size, saturation or quantity changed, see (12).

(12) *[W obu sutkach] [zagęszczenia], które były widoczne [w badaniu poprzednim z 1999r] i odpowiadają najpewniej [skupiskom resztkowej tkanki gruczołowej] ([w sutku lewym] jest nieco bardziej wysycone).*

[In both breasts] [densities] which were observed [in the previous report from 1999] and probably correspond to [glandular tissue concentration](the one [in the left breast] is slightly more saturated).

As the identified fragments do not have to be contiguous, we cannot unite them by applying a single grammar rule. Instead, we add a rule which recognizes the verb phrase in isolation, i.e., *obecnie nie jest widoczne* 'is no longer visible' or *jest nieco bardziej wysycone* 'is slightly more saturated', and then, we merge the results externally: if the verb is negated, the finding block is removed from the results and the structure referring to the previous examination is updated to indicate that the density is no longer present; if the finding undergone any other changes (size, quantity, saturation), only this information is updated.

As localization is also part of the finding block, it is deleted as well if the finding is no longer visible, and remains present otherwise. Therefore, the overgeneration of localization phrases, as in (11), is avoided.

7. Evaluation

To evaluate the results of incorporating the above changes, we tested the system on a set of the most problematic cases which were previously incorrectly annotated. Fig. 2 shows that, after the amendments, we recognized all but 4 findings, while only 2 were wrongly identified. Localization information was recognized quite well. We have introduced only two errors – both of them were caused by an atypical occurrence of the phrase *tej okolicy* 'this area'.

	before	now
patient records	27	27
FINDINGS	57	57
correctly recognized findings	40	53
including "relative" findings	–	16
unrecognized findings	17	4
badly recognized findings	2	2
correctly placed block beginnings	24	43
incorrectly placed block beginnings	16	10
incorrectly placed block endings	8	3
LOCALIZATION	130	139
incorrectly recognized	0	2
unrecognized	14	3
recognized "relative" localizations	–	3

Figure 2: Evaluation of automatically identifying blocks' boundaries/attributes on selected data

	nb	%
patient records	705	
FINDINGS	341	100
correctly recognized findings	321	94.13
unrecognized findings	20	5.86
badly recognized findings	34	9.97
correctly placed block beginnings	275	80.64
incorrectly placed block beginnings	46	13.49
incorrectly placed block endings	26	7.62
LOCALIZATION	2196	100
incorrectly recognized	4	0.18
unrecognized	25	1.14

Figure 3: Previous full evaluation on the test set of data

	nb	%
patient records	705	
FINDINGS	343	100
correctly recognized findings	334	97.37
including "relative" findings	13	3.74
unrecognized findings	9	2.62
badly recognized findings	34	9.91
correctly placed block beginnings	299	87.17
incorrectly placed block beginnings	35	10.20
incorrectly placed block endings	21	6.12
LOCALIZATION	2189	100
incorrectly recognized	3	0.14
unrecognized	20	0.91
recognized "relative" localizations	1	0.05

Figure 4: Current full evaluation on the test set of data

Figure 3 and 4 show[3] second comparison of the results obtained by two versions of the system: the one described

[3]The evaluation was done manually, what results in the slight difference between the numbers of findings.

in (Mykowiecka et al., 2005) and the one presented in this paper (with slight additional changes of the grammar and segmentation algorithms not addressed here). In this experiment we used a set of 705 randomly selected reports. The improvement of general results is less significant than in the case of the results obtained for the problematic cases but the overall goal of the work was achieved as the identification of findings improved a lot. In particular, the previously ignored 13 "relative" findings were recognized.

The impact of the system's modifications on recognition of localization phrases was less striking although the performance improved in this respect as well. It just turned out that there are many more ways to specify relative localisations than we have foreseen and the constructions described in the paper do not appear very often.

8. Conclusions

We presented a few simple techniques which allow us to deal with difficult natural language phenomena such as ellipsis, anaphora resolution, comparisons, coordination or negation. Our goal was to improve performance of the system processing medical data; providing general solutions to these hard problems requires deep processing and is far beyond the scope of this paper. Nevertheless, the adopted methods are fully sufficient for the described application, as indicated by the performance improvement.

9. Acknowledgements

We are grateful for the comments to two anonymous reviewers. The work presented in the paper was partly supported by KBN grant nr 3 T11C 007 27

10. References

Becker, M., W. Drożdżyński, H. Krieger, J.Piskorski, U. Schaefer, and F. Xu, 2002. SProUT — Shallow Processing with Typed Feature Structures and Unification. *In Proceedings of ICON 2002, Mumbai, India*.

Hahn, U., 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*:63–74.

Hripcsak, G., J. Austin, P. Anderson, and C. Friedman, 2002. Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports. *Radiology*, 224.

Hripcsak, G., C. Friedman, and P. Alderson, 1995. Unlocking clinical data from narrative reports. *Annals of Internal Medicine*, 122.

Jain, N. L. and Carol Friedman, 1997. 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*. pages 829–833.

Kupść, A., M. Marciniak, A. Mykowiecka, J. Piskorski, and T. Podsiadły-Marczynkowska, 2004. Information extraction from mammogram reports. In *Proceedings of KONVENS 2004*. Schriftenreihe der Osetrreischen Gesellschaft fur Artificial Intelligence.

Marciniak, M., A. Mykowiecka, A. Kupść, and J. Piskorski, 2004. Intelligent content extraction from Polish medical reports. In *Proceedings of ICMIT 2004*. PJWSTK.

Mykowiecka, A., A. Kupść, and M. Marciniak, 2005. Rule-based medical content extraction and classification. In *Proceedings of ISMIS 2005, Gdańsk*. Springer-Verlag.

Piskorski, J., 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.

Taira, Ricky K., Stephen G. Soderland, and Rex M. Jakobovits, 2001. Automatic structuring of radiology free-text reports. *Radiographics*, 21.