Rule-based Medical Content Extraction and Classification

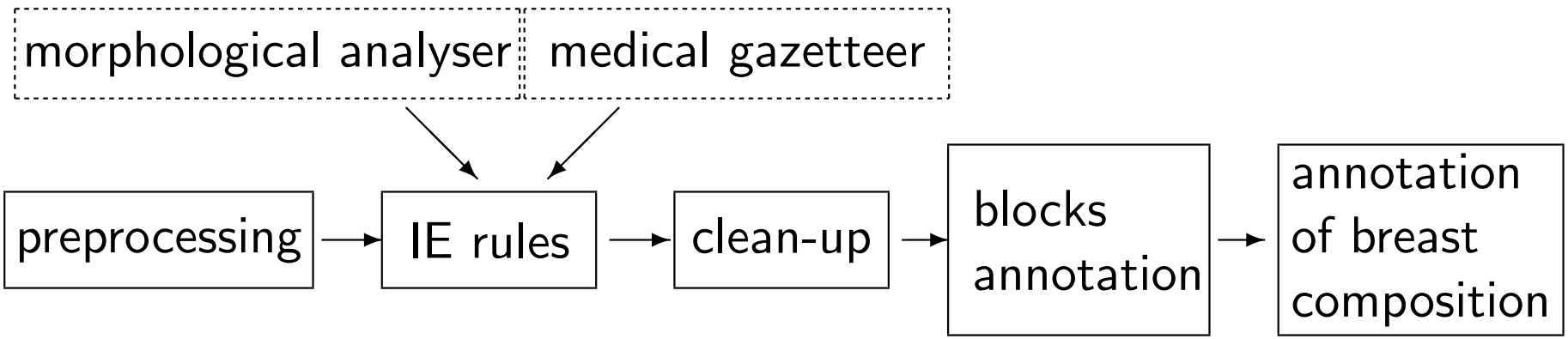
Małgorzata Marciniak, Agnieszka Mykowiecka

Task Description

- input data: about 2000 mammography reports from 3 Warsaw hospitals,
- output data: (for each report) very detailed information concerning:
  - breast composition,
  - anatomical findings observed and their description (shape, size, contour, etc.),
  - report summary.

Processing Strategy

- preprocessing: correction of spelling and punctuation,
- IE (SProUT) grammars extracting specific information (Polish morphology lexicon, domain lexicon, domain specific rules),
- postprocessing — data cleaning and information merging:
  - XML output cleaning,
  - block boundaries insertion,
  - segmentation of breast composition block,
- adding a report summary (classification).



A Sample Report

- Sutki o utkaniu gruczołowym. W sutku prawym w KGZ widoczne owalne, słabo wysycone, dobrze ograniczone zagęszczenie o wym. 15x10mm odpowiadające zmianie łagodnej – najpewniej torbieli. Zmian podejrzanych o złośliwość nie wykazano. Doły pachowe wolne. Kontrolna mammografia za rok.
- Breasts with glandular tissue. In the right breast in the upper outer quadrant there is an oval, low density, well circumscribed finding of 15x10mm. Benign finding – probably a cyst. No malignant findings were found. Armpits without any changes. A mammography checkup in a year.

up utp
log LOC|BODY_PART:breast||LOC|L_R:l-r *Breasts*
tsz BTISSUE:gll *with gladural tissue*
utk uk
zp
LOC|BODY_PART:breast || LOC|LOC_CONV:uoq || LOC|L_R:right
ANAT_CHANGE:density || CONTOUR:circ. || GRAM_MULT:singular
|| SATURATION:low || SHAPE:oval
DIM:mm || NUM1:15 || NUM2:10
DIAGNOSIS_RTG:benign
INTERPRETATION:cyst
zk
DIAG_RTG:no_susp || LOC|BODY_PART:breast || LOC|L_R:l-r
DIAG_RTG:no_susp || LOC|BODY_PART:armpit || LOC|L_R:l-r
RECOMMENDATION|FIRST:mmg||TIME|DESCRIPTOR:rok

MMG_REL:avg_reliable
REPORT_CLASS:diag_benign
REPORT_WITH_FINDINGS:yes

Attributes of Breasts' Composition

breasts' description	
BRSURG|SURGERY	undergone surgeries
BRSURG|REASON	the reason for the surgery
BTISSUE	main type of breast tissue
FIBRE	fibrosis
CHAR	features of breast tissue, e.g., displastic
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

Sample Phrases

- Sutki o przewadze tkanki tłuszczowej.
- Sutki z przewagą utkania tłuszczowego.
- Przewaga tkanki tłuszczowowej.
- Sutki o utkaniu z przewagą tłuszczowego.
- Sutki o budowie z przeważającą tkanką tłuszczową.
- ...

Breasts with predominant fat tissue.

A Sample Grammar Rule

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

Breasts' Composition Block

- Sutki o resztkowym utkaniu gruczołowym w kwadrantach górno-zewnętrznych. Przewaga tkanki tłuszczowej.
- Breasts with the remnant glandular tissue in upper-outer quadrants. Predominantly fat tissue.
- **up**
  LOC|BODY_PART:breast||LOC|L_R:left-right
  GLAND|QUANT:rem||BTISSUE:gll
  LOC|LOC_CONV:uoq
  BTISSUE:fat_gl
  uk

. . . after Processing

up
log LOC|BODY_PART:breast||LOC|L_R:left-right
utp
gland GLAND|QUANT:rem
tsz BTISSUE:gll
+ LOC|BODY_PART:breast||LOC|L_R:left-right
lsz LOC|LOC_CONV:uoq
utk
utp
+ LOC|BODY_PART:breast||LOC|L_R:left-right
tog BTISSUE:fat_gl
utk
uk

Another example

W sutku prawym w KGW 5mm dobrze ograniczone zagęszczenie (łagodne), drugie w KGZ o śr. 10mm równiez o podobnym charakterze.
In the right breast, in the upper inner quadrant, there is a well circumscribed density (benign), the second one in the uoq, of 10mm diameter size and a similar type.

```
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
    LOC|LOC_CONV:uoq
    DIM:mm||NUM1:10||NUM2:10
    TYPE_REF:yes
zk
```

Report Summary

- REPORT_CLASS — the summarised diagnosis. The value is inferred from component diagnosis and recommended examinations.
- MMG_REL — indicates how reliable the image is. The value depends on the type of the breast tissue and its features.
- REPORT_WITH_FINDINGS — a binary distinction specifying if any findings have been detected.

705 patient records	num	%
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
LOCALIZATIONS	2189	100
incorrectly recognized	3	0.14
unrecognized	20	0.91
BREAST COMPOSITION SUBBLOCKS	968	100.0
incorrectly recognized subblocks	9	1.0
unrecognized subblocks	3	0.3
incorrectly recognized positions of subblock endings	24	2.5

Conclusions

- The evaluation results:

	precision	recall
findings	90.76	97.38
block beginnings	81.25	97.07
localization	98.42	99.59
breasts' composition blocks	96.48	99.07

- The system is domain oriented (not portable).
- The extracted data is normalized and prepared to be inserted into a database for statistical analysis.

Thank you for your attention.