“Data-Driven” Ontologies for an Information Extraction System from Polish Mammography Reports

Agnieszka Mykowiecka¹, Małgorzata Marciniak¹, Teresa Podsiadły-Marczykowska²

¹IPI PAN Ordana 21, 01-237 Warsaw, Poland {agn,mm}@ipipan.waw.pl
²IBIB PAN Trojdena 4, 02-109 Warsaw, Poland teresa@ibib.waw.pl

10th International Protégé Conference July 15-18, 2007; Budapest, Hungary
Agenda

- Ontology - a method of knowledge representation for IE (Information Extraction) systems
- Reuse of existing resources
- BI-RADS based Mammographic Ontology
- Mammographic Report Ontology tailored for IE
- Mammography IE System and its evaluation
- Conclusions

10th International Protégé Conference July 15-18, 2007; Budapest, Hungary
Ontology - a method of knowledge representation for IE Systems

- Information extraction requires prior knowledge on data structures we would like to identify

- Information in mammography reports - composed and complicated - a theoretical approach of using the predefined domain knowledge is required
Reuse of existing resources

- Breast Cancer Image Ontology (BCIO) from MIAKT project
- NCI Cancer Ontology containing more than 17,000 concepts, but not mammography
- Basic Clinical Ontology for Breast Cancer from Stanford resources

no models suitable for reuse were found
too general, or covered related, but in fact distinct domain
BI-RADS based Mammographic Ontology (1)

Model is based on knowledge contained in BI-RADS, only extensions are concepts describing technical attributes of breast X-ray films mentioned in reports.
BI-RADS based Mammographic Ontology (2)

instances of class Lesion MMG form knowledge base of the model and are compared to masses description in authentic reports.
Mammographic Report Ontology tailored for IE (1)

- Why the need for the second model - after firsts IE experiments it was found that there is a discrepancy between mammographic terminology and the scope of general notions found in BI-RADS and those used in real life Polish radiology reports
- Second model (Mammographic Report Ontology) is needed extending the scope of the first model and its granularity
- Knowledge acquisition stage has been repeated
  - medical literature, additional reports, consultations with radiologists
- Main problems when developing Mammographic Report Ontology:
  - difficulties in delimiting a domain
  - difficulties with representing formal differences which are often neglected in real life texts
Mammographic Report Ontology tailored for IE (2)

- class **HumanAnatomy** - a part of human anatomy model
- class **Medicine** - containing informations related to mmg examination
- class **PhysicalFeature** - describing such physical features of mammographic lesions like shape, size, contour, density etc.
- class **Comparison** includes concepts used while comparing various types of features, e.g. number, level and size
- class **Time**

model adapted to needs of IE tools - enlarged scope of general notions
The IE application is implemented using the general system SProUT
• The IE application is implemented using the general system SProUT.

• For the purpose of being used inside the SProUT systems grammars, the ontology had to be translated into a Typed Feature Structures hierarchy.

• The class hierarchy is repeated as the TFS type hierarchy omitting only the highest level ontology classes which are outside the mammography domain.

• The properties are just attributes of type features structures used in SProUT.

• The main difference is introducing structures which combine elements of the ontology.

10th International Protégé Conference July 15-18, 2007; Budapest, Hungary
## Evaluation of IE System

<table>
<thead>
<tr>
<th>Type of information</th>
<th>precision</th>
<th>recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>pathological findings’ blocks beginnings</td>
<td>81,25</td>
<td>97,07</td>
</tr>
<tr>
<td>breasts’ composition blocks</td>
<td>96,48</td>
<td>99,07</td>
</tr>
<tr>
<td>pathological findings</td>
<td>92,44</td>
<td>97,46</td>
</tr>
<tr>
<td>pathological findings interpretation</td>
<td>98,19</td>
<td>93,69</td>
</tr>
<tr>
<td>all path. findings ( also those for which only interpretation was given )</td>
<td>90,76</td>
<td>97,38</td>
</tr>
<tr>
<td>localization</td>
<td>98,42</td>
<td>99,59</td>
</tr>
<tr>
<td>recommendation</td>
<td>98,63</td>
<td>99,5</td>
</tr>
</tbody>
</table>

Evaluation of a random set of 705 reports

10th International Protégé Conference July 15-18, 2007; Budapest, Hungary
Thank you
Sample Rule

wch_zm :

(morph & [POS noun, STEM "węzeł", INFL infl_noun &
[ NUMBER_NOUN #nb ] ] |
token & [SURFACE "ww"] |
gazetteer & [GTYPE gaz_med_wezel, G_CONCEPT lymph_node,
G_NUMBER #nb ] )

-> interpret_str & [INTERPRETATION intr_lymph_node,
MORPH agr & [N #nb]].
Mammography – a sample report

• 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).

• 775
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).
Mammography – Results

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

overall diagnosis

finding description

tissue block