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

Agnieszka Mykowiecka[1], Małgorzata Marciniak[1], Teresa Podsiadły-Marczykowska[2]
[1]IPI PAN Ordona 21, 01-237 Warsaw, Poland {agn,mm}@ipipan.waw.pl
[2] BIB PAN Trojdena 4, 02-109 Warsaw, Poland teresa@ibib.waw.pl

1. Introduction

The paper[1] describes the ontology development for an IE (Information Extraction) application for Polish mammography reports, experiences and lessons learned, and the evaluation of the system. Information extraction requires prior knowledge on data structures we would like to identify. When information being searched for is as complicated as this contained in mammography reports, a theoretical approach of using the predefined domain knowledge is required. For our research goal, extraction of possibly all precise information from mammography reports, ontology has been chosen as platform of knowledge representation. During the work two ontologies have been developed, the first based mainly on BI-RADS [5], the second adjusted to the task of information extraction. The paper is structured as follows: section 2 relates our experiences concerning the reuse of existing ontologies, sections 3 and 4 present respectively initial mammographic ontology and modified model adapted for the task of information extraction, section 5 presents the IE system and results of information extractions, section 6 concludes the paper.

2. Available Domain Ontologies

It is a well known fact, that developing ontology from scratch is a time and work consuming enterprise. That's why at the beginning of our work we reviewed repositories of biomedical ontologies in the hope that we could reuse them with some minor modifications. The most suitable seemed Breast Cancer Image Ontology (BCIO) from MIAKT project [1], but this ontology is not publicly available. Paper [2] in which this ontology is reported gives only some details on the model structure and design decisions. Large NCI Cancer Ontology contains more than 17 000 concepts, but not mammography [3]. Basic Clinical Ontology for Breast Cancer [4] from Stanford resources is a model directly tied to mammography domain, but all the same not exactly appropriate for our purpose because it contains notions used by physician to describe clinical state of breast and auxiliary lymph nodes, and also concepts witch can be used in epidemiological breast cancer study, but no notions necessary for breast X-ray films description and interpretation. To sum up: no models suitable for reuse were found. They were either not publicly available or covered related, but in fact distinct domains. To sum up: no models suitable for reuse were found, they were either too general or covered related, but in fact distinct domain. All this rendered them of little help for us.

3. First model – Bi-RADS based Mammographic Ontology

BI-RADS (Breast Imaging Reporting and Database System), standardized terminological system in mammography, has been used as a starting point in knowledge acquisition for initial mammographic ontology development. It contains rudimentary lexicon allowing the description of basic visual features of masses and calcifications and some diagnostic criterions. First mammographic ontology has been developed basing mainly on knowledge contained in BI-RADS, only extensions are concepts describing technical attributes of breast X-ray films mentioned in reports (see Fig. 1 panel A). Lesion scope and attributes correspond to lesions mentioned in BI-RADS. Every class representing mammographic lesion contains as its subclass a class whose instances represent notions described in reports. For example instances of classes ReportedMass form knowledge base of the model and are compared to masses description in authentic reports (see Fig.1 panel B). The model has been is reported in [6]. The ontology has been implemented in frame Protégé ver. 2.1.1. Now it is migrated to OWL using Protégé 3.3.

4. Second model - Mammographic Report Ontology Tailored for IE application

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. To improve performance of IE system, we decided to build the second model (mammographic report ontology) by extending the scope of the first model and its granularity. Knowledge acquisition stage (including

[1] This work was financed by the Polish national KBN project number 3 T11C 007 27.

additional sources such as: analysis of medical literature, analysis of additional corpus reports, and finally consultations with radiologists concerning interpretation and description of mammograms) has been repeated. The resulting ontology contains following main parts (see Fig.1):

- a fragment of human anatomy description (*AnatomicalPathology, HumanBodyParts* and *HumanTissue*),
- *Medicine* type containing a proper mammography information part being the *MedExamination* subtype and related medical information under the types of *MedProcedure*, *MedJudgement*, *MedExamOrProcReason* and *MedRecommendation*.

In the domain of mammography reports, some very general concepts are also used. The *PhysicalFeature* class describes such physical features like shape, size, contour, density etc. The *Comparison* class includes concepts used while comparing various types of features, e.g. number, level and size. The external *Time* ontology was planned to be used but as time information here is very simple (only dates and periods of time in months or years) we also defined simple *Time* module which can be replaced by something more elaborated in the future.

The problems of defining the mammography ontology were of two kinds:

- difficulties in representing different concepts which are not always distinguished in real life texts
- difficulties in delimiting a domain.

The main problem, which occurred after reviewing some of the reports, was the lack of clear differentiation between what is visible on the mammogram and what lesions interpretation is. If a doctor is pretty sure what the nature of a finding is, he/she writes only an interpretation. However, as the difference between what is seen and its interpretation is important, we represent it in the ontology and the problem is solved inside the IE system. The next problem to be solved is which information should be treated as a domain internal and which should be represented outside. For general concepts like shape or size, the first solution would result in repeating the same value in many domains. In the presented ontology, we adopted the idea of identifying general concepts and represent them outside the particular domain. In this case however, problem of identifying subsets of valid values for a given general type arises. It can be solved by defining the appropriate restrictions. For the purpose of mammography ontology, we defined only the values needed within the domain. For example, in the *PhysicalFeature* class representing 2-dimmensional objects' shapes, only these which actually occur in the mammography domain are defined.

PANEL A **PANEL B** **PANEL C**

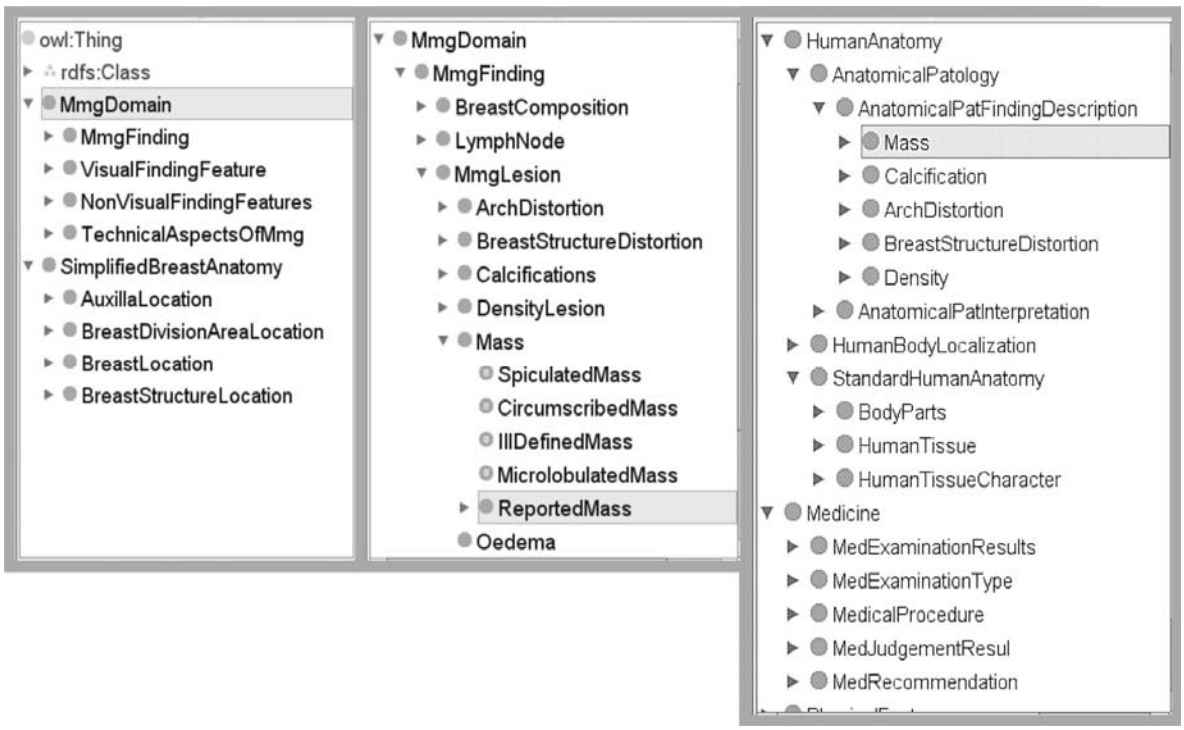

Fig.1 Two models demonstrative comparison - scope and granularity. Panels A and B presents the first, Bi-RADS based model of mammography. Panel A – general concepts. Panel B – lesion types. Panel C – model adapted to needs of information extraction tools, main difference seen on screen shot is enlarged scope of general notions.

5. Information Extraction System and its Evaluation

The IE application is implemented using the general system SProUT [7]. For the purpose of being used inside the SProUT systems grammars, the ontology had to be translated into a typed feature hierarchy. The correspondence of the hierarchies is quite high. The class hierarchy is repeated as the type hierarchy omitting only the high 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. They are used in rules for recognizing typical phrases which combine information of different kinds, e.g. mixed conventional and anatomical localizations, localization together with diagnosis. As IE rules can produce only one structure, introducing such multi rooted structures were necessary.

The mammography system [8] transforms texts into a set of typed feature structures. Besides using SProUT shallow grammars, several post-processing Perl scripts were developed. They remove duplicate analyses, delete irrelevant information, and aggregate the extracted data according to the domain model. The process of the information aggregation is difficult as a single report can contain several mammographic findings. In order to create their separate descriptions, we have implemented several heuristics, which group extracted attribute-value pairs into a consistent description of a finding.

The evaluation of the system was done on 705 new mammogram reports. The results 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 problem observed was the improper recognition of the beginning and the end of information blocks. Some errors were caused by incomplete grammar coverage especially for negation and comparison phenomena and words with different meaning depending on a context.

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

Fig. 2 Evaluation of a random set of 705 reports

6. Conclusions

The initial, based on standardized mammographic lexicon ontology, turned out to be insufficient for the task of information extraction from mammography reports. Those brief and compact texts enclose in fact knowledge from some other domains than mammography alone, and that knowledge is more detailed than controlled dictionaries. Necessary modifications of the first model included granularity and scope of concepts, structure of the model and modeling patterns, and finally adaptation to specific IE system requirements. Those adjustments resulted in better IE system performance.

References

1. Breast Cancer Image Ontology (BCIO) http://www.aktors.org/miakt , last visited 23.04.2007 not publicly available, authorized access required.
2. Dasmahapatra S., Dupplaw D., Bo Hu, Lewis H., Lewis P., Shadbolt N.: Facilitating Multi-Disciplinary Knowledge-Based Suport for Breast Cancer Screening, Int. J. of Healthacare Technology and Menagement, vol. 7, no. 5/2006 pp 403-420
3. NCI Cancer Ontology http://swserwer.cs.vu.nl/partitioning/NCI/ last visited 23.04.2007
4. Basic Clinical Ontology for Breast Cancer http://acl.icnet.uk/~mw/ last visited 23.04.2007
5. Kopans D.B. et all. Breast Imaging Reporting and database System (BI-RADS), ACR 1997
6. Podsiadły-Marczykowska T., Guzik A., Mammographic Ontology − Conceptual Model of the Domain, The International Journal of Artificial Organs Vol. l27, no. 7, 2004
7. Drożdżyński W., Krieger H-U., Piskorski J., Schäfer U., Xu F., Shallow Processing with Unification and Typed Feature Structures − Foundations and Applications, *German AI Journal KI-Zeitschrif* , 2004.
8. Mykowiecka A., Kupść A., Marciniak M., Rule-based Medical Content Extraction and Classification, *Proceedings of ISMIS 2005*, Springer-Verlag, 2005.