Information Extraction from Patients' Free Form Documentation

Agnieszka Mykowiecka
Institute of Computer Science, PAS
Ordona 21, 01-237 Warszawa, Poland
agn@ipipan.waw.pl

Małgorzata Marciniak
Institute of Computer Science, PAS
Ordona 21, 01-237 Warszawa, Poland
mm@ipipan.waw.pl

Abstract

The paper presents two rule-based information extraction (IE) from two types of patients' documentation in Polish. For both document types, values of sets of attributes were assigned using specially designed grammars.

1 Method/General Assumptions

Various rule-based, statistical, and machine learning methods have been developed for the purpose of information extraction. Unfortunately, they have rarely been tested on Polish texts, whose rich inflectional morphology and relatively free word order is challenging. Here, we present results of two experiments aimed at extracting information from mammography reports and hospital records of diabetic patients.[1] Since there are no annotated corpora of Polish medical text which can be used in supervised statistical methods, and we do not have enough data for weakly supervised methods, we chose the rule-based extraction schema. The processing procedure in both experiments consisted of four stages: text preprocessing, application of IE rules based on the morphological information and domain lexicons, postprocessing (data cleaning and structuring), and conversion into a relational database.

Preprocessing included format unification, data anonymization, and (for mammography reports) automatic spelling correction.

The extraction rules were defined as grammars of the SProUT system, (Drożdżyński et al., 2004). SProUT consists of a set of processing components for basic linguistic operations, including tokenization, sentence splitting, morphological analysis (for Polish we use Morfeusz (Woliński, 2006)) and gazetteer lookup. The SproUT components are combined into a pipeline that generates typed feature structures (TFS), on which rules in the form of regular expressions with unification can operate. Small specialized lexicons containing both morphological and semantic (concept names) information have been created for both document types.

Extracted attribute values are stored in a relational database.[2] Before that, mammography reports results undergo additional postprocessing — grouping together of extracted data. Specially designed scripts put limits that separate descriptions of anatomical changes, tissue structure, and diagnosis. More details about mammography IE system can be found in (Mykowiecka et al., 2005).

2 Document types

For both document types, partial ontologies were defined on the basis of sample data and expert knowledge. To formalize them, we used OWL-DL standard and the *Protégé* ontology editor. The excerpt from the ontology is presented in Fig. 1.

In both cases, the relevant part of the ontology was translated into a TFS hierarchy. This resulted in 176 types with 66 attributes for the mammography domain, and 139 types (including 75 drug names) with 65 attributes for diabetic patients' records.

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

[2] This last stage is completed for the diabetes reports while for mammography it is still under development.

BioNLP 2007: Biological, translational, and clinical language processing, pages 181–182,
Prague, June 2007.

BiochemicalData: BloodData: HB1C
Diet
DiseaseOrSymptom
  Disease
    AutoimmuneDisease
    Cancer
    Diabetes: Type1, Type2, TypeOther
  Symptom
    Angiopathy: Macroangiopathy, Microangiopathy
    BoodSymptom: Hypoglicaemia
    Neuropathy: Autonomic, PeripheralPolineuropathy
    UrineSymptom: Acetonuria, Microalbuminuria
Medicine
  DiabeticMedicine: Insulin, OralDiabeticMedicine
AnatomicalLocalization
  BodyPart
    Breast: Subareola, urq, ulq, lrq, llq
  BodySide: Left, Right
HistDiagnosis: Benign, Suspicious, Malignant
TissueSpecification: GlandularTissue, FatTissue

Figure 1: A sample of classes

3 Extraction Grammars

The number of rules is highly related to the number of attributes and possible ways of formulating their values. The grammar for mammography reports contains 190 rules; that for hospital records contains about 100 rules. For the first task, nearly the entire text is covered by the rules, while for the second, only a small part of the text is extracted (e.g., from many blood tests we are interested only in HBA1C). Polish inflection is handled by using the morphological analyzer and by inserting the most frequent morphological forms into the gazetteer. Free word order is handled either by rules which describe all possible orderings, or by extracting small pieces of information which are merged at the postprocessing stage. Fig. 2 presents a fragment of one mammography note and its output. The *zp* and *zk* markers are inserted during the information structuring stage to represent borders of an anatomical change description. Similar markers are introduced to structure the tissue description part.

4 Evaluation

The experiments were evaluated on a set of previously unseen reports. Extraction of the following structures was evaluated: 1) simple attributes (e.g. diabetes balance); 2) structured attributes (e.g. localization); and 3) complex structures (e.g. description of abnormal findings). Evaluation of three selected attributes from both sets is given in Fig. 3. The worse results for unbalanced diabetes recognition were due to an unpredicted expression type.

W obu sutkach rozsiane pojedyncze mikrozwapnienia o charakterze łagodnym. Doły pachowe prawidłowe. Kontrolna mammografia za rok.
(*Within both breasts there are singular benign microcalcifications. Armpits normal. Next control mammography in a year.*)

zp LOC|BODY_PART:breast||LOC|L_R:left-right
   ANAT_CHANGE:micro||GRAM_MULT:plural
zk DIAGNOSIS_RTG:benign
   DIAGNOSIS_RTG:no_susp||LOC_D|BODY_PART:
     armpit||LOC_D|L_R:left-right
   RECOMMENDATION|FIRST:mmg||TIME:year

Figure 2: A fragment of an annotated mammography report

mammography – 705 reports			
	cases	precision	recall
findings	343	90.76	97.38
block beginnings	299	81.25	97.07
localizations	2189	98.42	99.59
diabetes – 99 reports			
unbalanced diabetes	58	96,67	69,05
diabetic education	39	97,50	97,50
neuropathy	30	100	96,77

Figure 3: Evaluation results for selected attributes

5 Conclusions

Despite the fact that rule based extraction is typically seen as too time consuming, we claim that in the case of very detailed information searching, designing rules on the basis of expert knowledge is in fact a method of a real practical value. In the next stage, we plan to use our tools for creating annotated corpora of medical texts (manually corrected). These data can be used to train statistical IE models and to evaluate other extraction systems.

References

Agnieszka Mykowiecka, Anna Kupść, Małgorzata Marciniak. 2005. Rule-based Medical Content Extraction and Classification, *Proc. of IIS: IIPWM05*. Advances in Soft Comp., Vol. 31, Springer-Verlag.

Witold Drożdżyński and Hans-Ulrich Krieger and Jakub Piskorski and Ulrich Schäfer and Feiyu Xu. 2004. Shallow Processing with Unification and Typed Feature Structures – Foundations and Applications. *German AI Journal KI-Zeitschrift, 01/04*.

Marcin Woliński. 2006. Morfeusz – a Practical Tool for the Morphological Analysis of Polish, *Proc. of IIS: IIPWM06*. Adv. in Soft Comp., Springer-Verlag.