Data-induced Domain Models for Information Extraction from Patients' Clinical Data

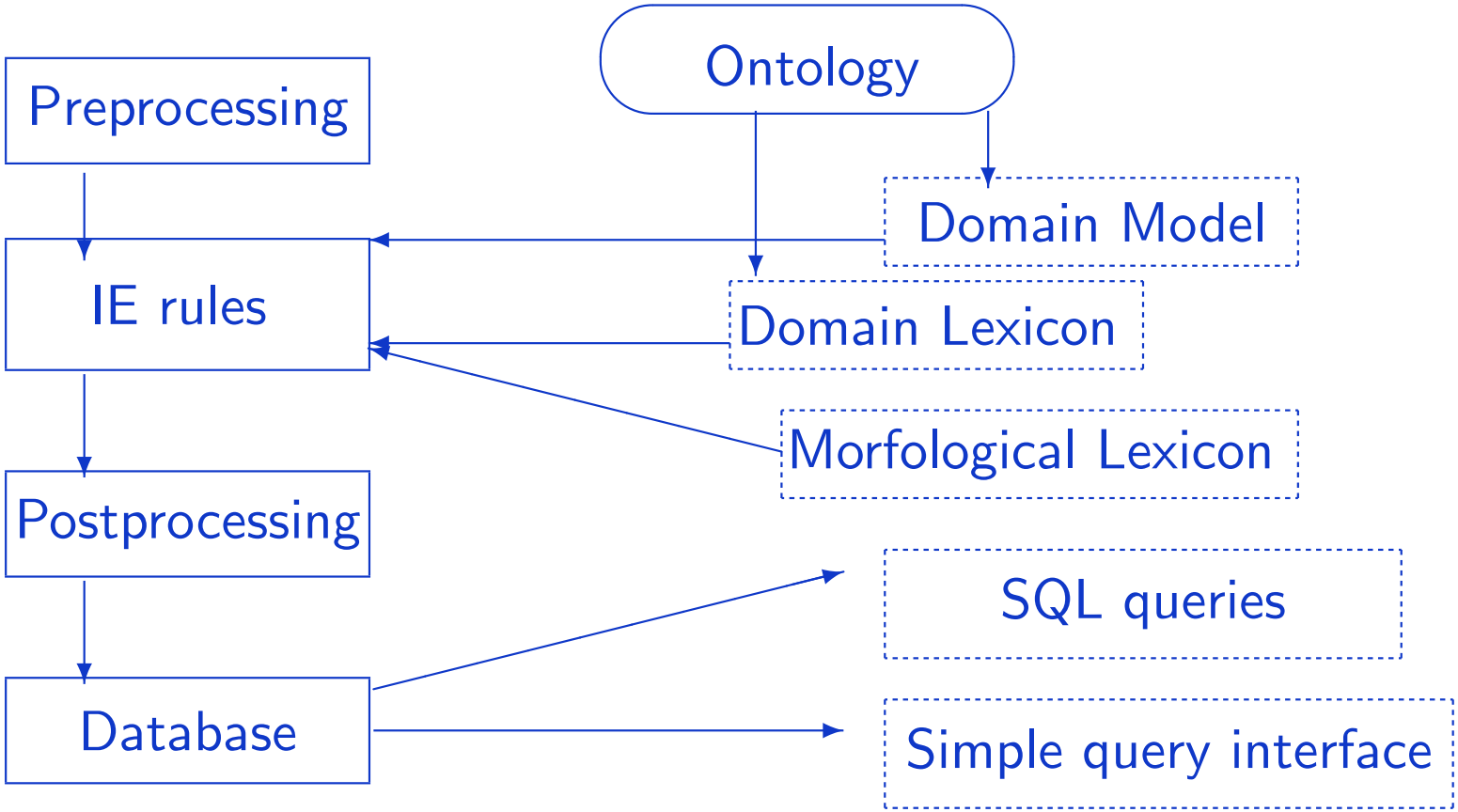
Agnieszka Mykowiecka and Małgorzata Marciniak

Institute of Computer Science, Polish Academy of Sciences, Warszawa, Poland
agn@ipipan.waw.pl, mm@ipipan.waw.pl

Overview

- **rule-based medical information extraction** (IE) for **Polish** medical texts using specially defined domain ontologies (first such an experiment carried out for Polish – a highly inflected and relatively free-word order language),
- two IE applications designed to select data from patients' documentation (**mammogram notes and diabetic patients' hospital records**) in Polish (unrestricted free texts),
- for most features precision and recall well above 80% were obtained.

Processing Stages



Ontology

- The ontology covers only the chosen medicine subdomains but it is coherent and can be further extended. At the moment only high level concepts are shared between domains.
- Some fragments of more general ontologies have been also defined: *PhysicalFeature*, *PhysicalFeatureComparison* and *Time*. They cover such physical features as size, contour, aggregation, density, projection and regularity and comparisons of quantity, degree and level.
- *Time* ontology covers only those cases which occur in selected type of documents. This means period of time in years, months and weeks, precise and imprecise dates, and also repetitions like *every year*.
- The ontology was manually translated into a TFS hierarchy. This resulted in 176 types with 66 attributes for the mammography domain and 139 types with 65 attributes for the diabetes.

IE-rule

An exemplary (simplified) rule cited below states that there are no long-lasting complications of diabetes of a certain type (identified as a gazetteer entry denoting a complication — variable #t).

```
brak_powiklan:> morph & [STEM "nie"]
(morph & [STEM "stwierdzić"] | morph & [STEM "wykryć"])
(morph & [STEM "obecność"])? morph & [STEM "późny"]
(morph & [STEM "cukrzycowy"] | morph & [STEM "cukrzyca"])
(morph & [STEM "w"] | morph & [STEM "pod"] )
(morph & [STEM "postać"] | morph & [STEM "typ"] ])
gazetteer & [GTYPE gaz_comp, G_CONCEPT #t]
-> no_comp_str & [N_COMP #t].
```

Selected results

Evaluation results for selected attributes for the 705 mammography reports:

	cases	precision	recall
findings	343	90.76	97.38
block beginnings	299	81.25	97.07
localizations	2189	98.42	99.59

Evaluation results for selected attributes for the 100 diabetes reports:

	cases	precision	recall
unbalanced diabetes	58	96,67	69,05
diabetic education	39	97,50	97,50
neuropathy	30	100	96,77

Conclusions

- rule-based IE applications using domain ontologies and linguistic data are capable for extracting very detailed specific information from texts in natural languages,
- for the selected task of IE from patients' clinical data, reusing existing ontologies was impossible due to the lack of appropriate resources,
- for methods applying NLP techniques, restricted simplified domain ontologies are necessary – good solution would be to enhance ontologies with a concept of views,
- for processing of languages with rich inflection patterns, ontologies should be linked with inflectional dictionaries.