Ekstrakcja informacji złożonych z tekstów w języku polskim

Agnieszka Mykowiecka i Małgorzata Marciniak

Projekt: *Opracowanie narzędzi do ekstrakcji informacji z tekstów w języku polskim*
wykonawcy m.in. Anna Kupść i Jakub Piskorski

26 lutego 2008 roku

Project Motivations and Goals

Motivations

- not many efforts on IE on Polish texts in contrast to many existing applications for many languages,
- existing IE tools couldnot be directly used for processing Polish.

Goals

- adapting chosen IE tools for processing Polish,
- collecting some linguistic resources for IE.

Activities

- adapting IE platforms SPoUT and (recently) GATE for tokenization and morphological analysis of Polish texts;
- collecting resourses and IE grammars for named entities recognition (NER) in Polish texts,
- ruled based IE experiments in a selected domain (medical texts),
- testing methods of terminology extraction on Polish data.

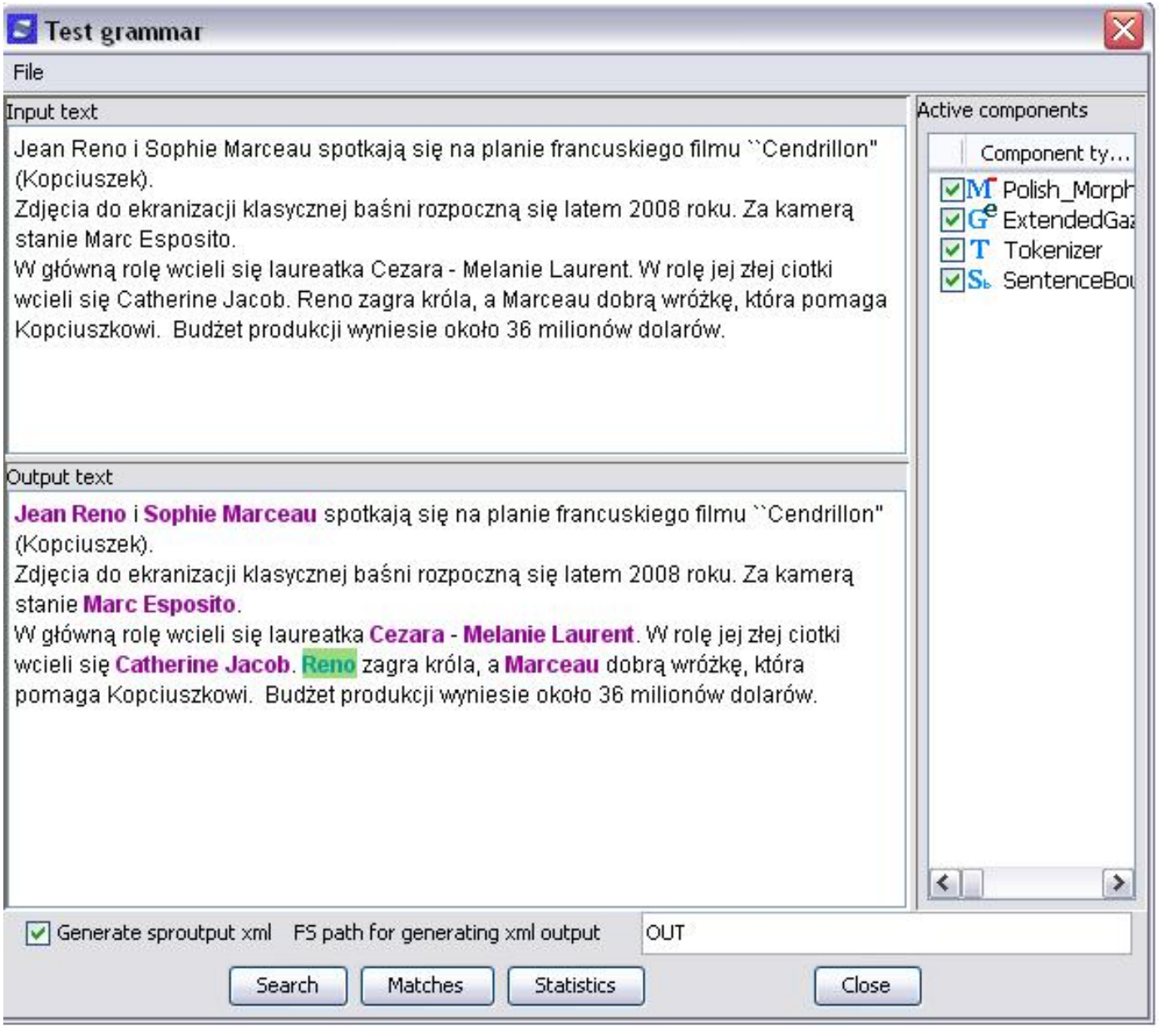
Talk Overview

- named entites – resources and IE grammars,
- rule based medical IE system,
- terminology extraction,
- concluding remarks.

NER Task Motivation

- Proper names are frequent: $\sim$ 10% of (newspaper) texts,
- Important for automatic text understanding (IE, QA, summarization . . .)
- Main problems:
  - Coreference resolution (*Jean Reno, Reno*),
  - Lemmatization for languages with rich inflection.

Jean Reno i Sophie Marceau spotkają się na planie francuskiego filmu "Cendrillon" ("Kopciuszek"). Zdjęcia do ekranizacji klasycznej baśni rozpoczną się latem 2008 roku. Za kamerą stanie Marc Esposito. W główną rolę wcieli się laureatka Cezara - Melanie Laurent. W rolę jej złej ciotki wcieli się Catherine Jacob. Reno zagra króla, a Marceau dobrą wróżkę, która pomaga Kopciuszkowi. Budżet produkcji wyniesie około 36 milionów dolarów.


Test grammar
File
Input text
Jean Reno i Sophie Marceau spotkają się na planie francuskiego filmu ``Cendrillon" (Kopciuszek).
Zdjęcia do ekranizacji klasycznej baśni rozpoczną się latem 2008 roku. Za kamerą stanie Marc Esposito.
W główną rolę wcieli się laureatka Cezara - Melanie Laurent. W rolę jej złej ciotki wcieli się Catherine Jacob. Reno zagra króla, a Marceau dobrą wróżkę, która pomaga Kopciuszkowi. Budżet produkcji wyniesie około 36 milionów dolarów.
Active components
Component ty...
Polish_Morph
ExtendedGaz
Tokenizer
SentenceBou
Output text
Jean Reno i Sophie Marceau spotkają się na planie francuskiego filmu ``Cendrillon" (Kopciuszek).
Zdjęcia do ekranizacji klasycznej baśni rozpoczną się latem 2008 roku. Za kamerą stanie Marc Esposito.
W główną rolę wcieli się laureatka Cezara - Melanie Laurent. W rolę jej złej ciotki wcieli się Catherine Jacob. Reno zagra króla, a Marceau dobrą wróżkę, która pomaga Kopciuszkowi. Budżet produkcji wyniesie około 36 milionów dolarów.
Generate sproutput xml
FS path for generating xml output
OUT
Search
Matches
Statistics
Close

Person Identifying Phrases

- *dr Jan Kowalski, dr. Jana Kowalskiego, prof. dr inż Paweł Zimny,*
- *Szef Kancelarii Premiera Mariusz Błaszczak,*
  Dyrektor Kancelarii Prezydenta Miasta Magdalena Samora,
- *Rektor Uniwersytetu Warszawskiego prof. dr hab. Włodzimierz Siwiński*
- *Aldona Wojtczak, dyrektor zarządzający (executive director)*
- *Jan K, Jan Maria Rokita,*
- *Halina Kowalska-Nowak,*
- *John J. Duncan, Jr.,*
- *[film] Altmana (Altman's film),*
- *[zdaniem/według] Hun Sena (as ... said /according to).*

General Assumptions

- NER as an information extraction task,
- rules based on lexicons and contexts heuristics,
- using SProUT system (Shallow Processing with Unification and Typed Feature Structures, DFKI), typical IE system enhanced with unification, to express IE rules which we enriched with Morfeusz morphological analyzer,
- information sources:
  - tokenizer,
  - sentence spliter,
  - general lexicon (used only for very small group of words),
  - specialized lexicon (gazetteer):
    - first names,
    - positions, titles, infixes, like 'van', 'van der',
    - some nouns and verbs – name preceeding keywords.

Fragment of the Domain Lexicon

Anna | GTYPE:gaz_given_name | G_CONCEPT:Anna |
G_GENDER:fem | G_CASE:nom_voc | G_BLOCKLETTERS:false
Anno | GTYPE:gaz_given_name | G_CONCEPT:Anna |
G_GENDER:fem | G_CASE:voc | G_BLOCKLETTERS:false
Annie | GTYPE:gaz_given_name | G_CONCEPT:Anna |
G_GENDER:fem | G_CASE:dat_loc | G_BLOCKLETTERS:false
Dyrektor | GTYPE:gaz_position | G_CONCEPT:dyrektor |
G_CASE:nom | G_NUMBER:singular
dr | GTYPE:gaz_title
dr. med. | GTYPE:gaz_title

NE Identification rules – an example

;; rule for person names including position, and missing first names
;; eg. *Dyrektor Scherr-Kowalski, Jr., Dyrektor A. Kowalski*

```
pl_person_2 :/       (((@seek(full_position) & #position)
                        ((token & [TYPE comma]) ?)
                       (@seek(title) & #title) ? ) |
                      (((@seek(full_position) & #position)
                       (token & [TYPE comma]) ?) ?
                        (@seek(title) & #title)) )
                      (@seek(last_name) & #last_name)
                      (@seek(name_suffix) & #suffix) ?
-> ne-person & [TITLE #title,
                      SURNAME #final_last_name,
                      P-POSITION #position,
                      NAME-SUFFIX #suffix],
where #final_last_name = LemmatizeLastNamePL(#last_name).
```

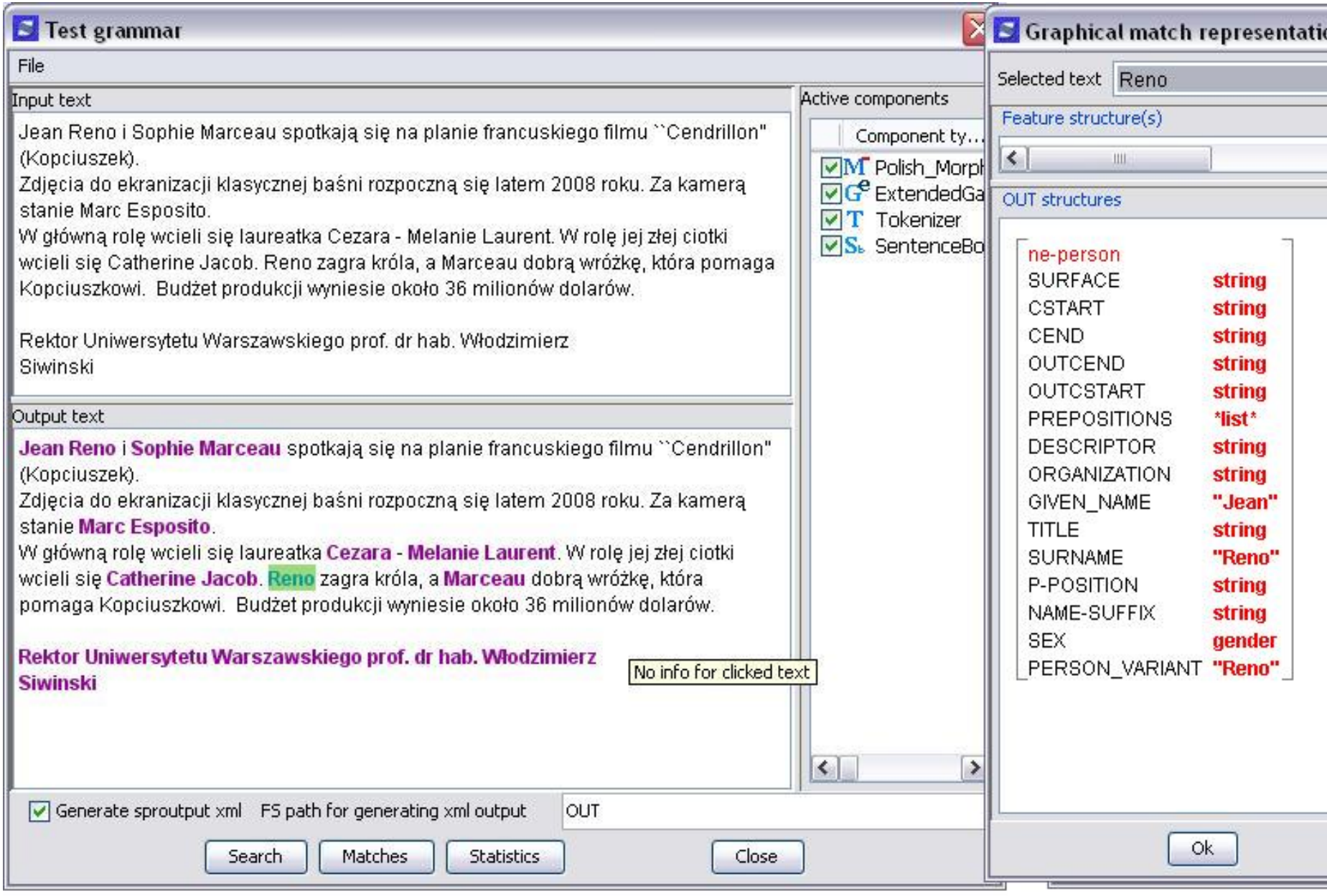
Exemplary "less credible" Rule

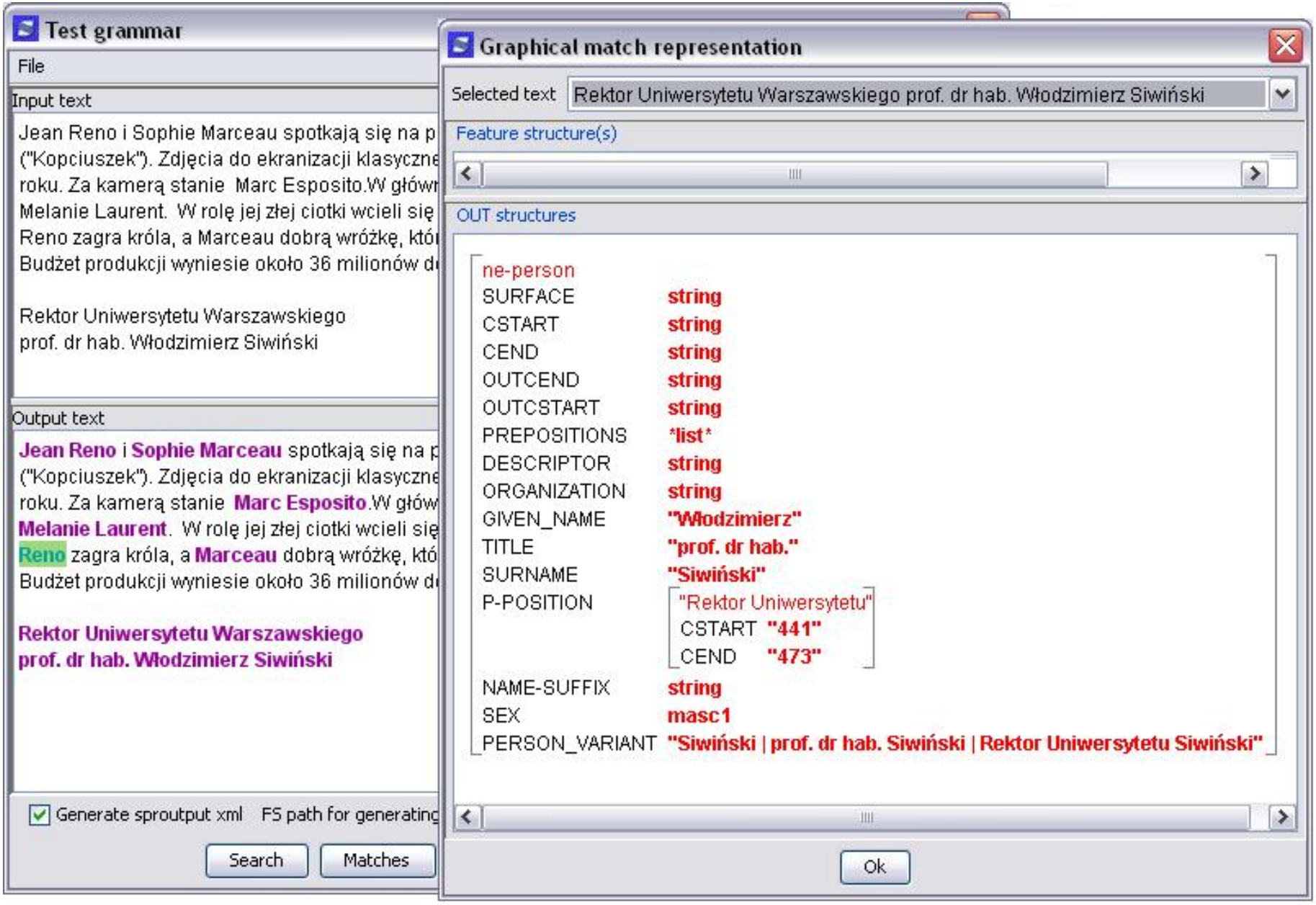
rule taking into account preceding nouns which are often followed by a last name (list of such nouns is encoded directly in the type hierarchy, eg. *film Formana*,

```
pl_person_preceded_by_noun :>
    (morph &                    [STEM noun_preceding_person_name,
                                 SURFACE #word])
     (@seek(last_name) &        [N_SURFACE #last_name,
                                 CSTART #s, CEND #e])

-> ne-person &                  [OUTCSTART #s, OUTCEND #e,
                                 SURNAME #last_name],
                                 where NotCapitalized(#word).
```

noun_preceding_person_name: plan, teoria, film, opera ...
verb_preceding_person_name: powiedział, podsumował ...


Test grammar
File
Input text
Jean Reno i Sophie Marceau spotkają się na planie francuskiego filmu ``Cendrillon" (Kopciuszek).
Zdjęcia do ekranizacji klasycznej baśni rozpoczną się latem 2008 roku. Za kamerą stanie Marc Esposito.
W główną rolę wcieli się laureatka Cezara - Melanie Laurent. W rolę jej złej ciotki wcieli się Catherine Jacob. Reno zagra króla, a Marceau dobrą wróżkę, która pomaga Kopciuszkowi. Budżet produkcji wyniesie około 36 milionów dolarów.
Rektor Uniwersytetu Warszawskiego prof. dr hab. Włodzimierz Siwinski
Output text
Jean Reno i Sophie Marceau spotkają się na planie francuskiego filmu ``Cendrillon" (Kopciuszek).
Zdjęcia do ekranizacji klasycznej baśni rozpoczną się latem 2008 roku. Za kamerą stanie Marc Esposito.
W główną rolę wcieli się laureatka Cezara - Melanie Laurent. W rolę jej złej ciotki wcieli się Catherine Jacob. Reno zagra króla, a Marceau dobrą wróżkę, która pomaga Kopciuszkowi. Budżet produkcji wyniesie około 36 milionów dolarów.
Rektor Uniwersytetu Warszawskiego prof. dr hab. Włodzimierz Siwinski
No info for clicked text
Active components
Component ty...
Polish_Morph
ExtendedGa
Tokenizer
SentenceBo
Generate sproutput xml
FS path for generating xml output
OUT
Search
Matches
Statistics
Close
Graphical match representatio
Selected text
Reno
Feature structure(s)
OUT structures
ne-person
SURFACE string
CSTART string
CEND string
OUTCEND string
OUTCSTART string
PREPOSITIONS *list*
DESCRIPTOR string
ORGANIZATION string
GIVEN_NAME "Jean"
TITLE string
SURNAME "Reno"
P-POSITION string
NAME-SUFFIX string
SEX gender
PERSON_VARIANT "Reno"
Ok


Test grammar
File
Input text
Jean Reno i Sophie Marceau spotkają się na p
("Kopciuszek"). Zdjęcia do ekranizacji klasyczne
roku. Za kamerą stanie Marc Esposito.W główn
Melanie Laurent. W rolę jej złej ciotki wcieli się
Reno zagra króla, a Marceau dobrą wróżkę, któ
Budżet produkcji wyniesie około 36 milionów d
Rektor Uniwersytetu Warszawskiego
prof. dr hab. Włodzimierz Siwiński
Output text
Jean Reno i Sophie Marceau spotkają się na p
("Kopciuszek"). Zdjęcia do ekranizacji klasyczne
roku. Za kamerą stanie Marc Esposito.W głów
Melanie Laurent. W rolę jej złej ciotki wcieli się
Reno zagra króla, a Marceau dobrą wróżkę, któ
Budżet produkcji wyniesie około 36 milionów d
Rektor Uniwersytetu Warszawskiego
prof. dr hab. Włodzimierz Siwiński
Generate sproutput xml
FS path for generating
Search
Matches
Graphical match representation
Selected text
Rektor Uniwersytetu Warszawskiego prof. dr hab. Włodzimierz Siwiński
Feature structure(s)
OUT structures
ne-person
SURFACE string
CSTART string
CEND string
OUTCEND string
OUTCSTART string
PREPOSITIONS *list*
DESCRIPTOR string
ORGANIZATION string
GIVEN_NAME "Włodzimierz"
TITLE "prof. dr hab."
SURNAME "Siwiński"
P-POSITION "Rektor Uniwersytetu"
CSTART "441"
CEND "473"
NAME-SUFFIX string
SEX masc1
PERSON_VARIANT "Siwiński | prof. dr hab. Siwiński | Rektor Uniwersytetu Siwiński"
Ok

Inflection of Person Names

- Category (POS) and gender of the surname

  feminine noun: *belka* 'beam'
  nom: Marek Belka (m) vs. Maria Belka (f)
  gen: Marka Belki (m) vs. Marii Belki (f)
  masculine noun: *Polak* 'Pole'
  nom: Kazimierz Polak (m) vs. Kazimiera Polak (f)
  gen: Kazimierza Polaka (m) vs. Kazimiery Polak (f)
  adjective: *niski* 'Short'
  nom: Kazimierz Niski (m) vs. Kazimiera Niska (f)
  gen: Kazimierza Niskiego vs. Kazimiery Niskiej

- Origin and pronunciation of the name
  nom: Oscar Wilde (E) vs. Hans Wilde (G)
  gen: Oscara Wilde'a (E) vs. Hansa Wildego (G)

Lemmatization

There are three categories of rules for finding base forms defined:

- rules for typical Polish surnames return basic form of the first name form the lexicon + standard change of the last name's suffix:
  - *z dyrektorem Karoliną Kowalską* → *dyrektor Karolina Kowalska*,
  - *o rektorze Janie Kowalskim* → *rektor Jan Kowalski*
- rules based on morphology return basic form of the first name form the lexicon + basic form of the second name from the dictionary if is exists there, eg. *Jana Kruka*→ *Jan Kruk*,
- relaxed rules based on suffixes and token types,

Statistical Lemmatization

Motivation

- Help for rule-based lemmatization: document- and collection-level name matching
- Test portability and extend metrics applied to English name matching (Christen, 2006)

Different string distance metrics tested:

- single token: edit distance metrics, eg. *Levenstein* (edit transformations), character-level *q*-grams, common prefix,
- multitoken distance metrics: *Longest common substrings*, *Weighted LCS*, *Jaro* and *Jaro-Winkler* ...

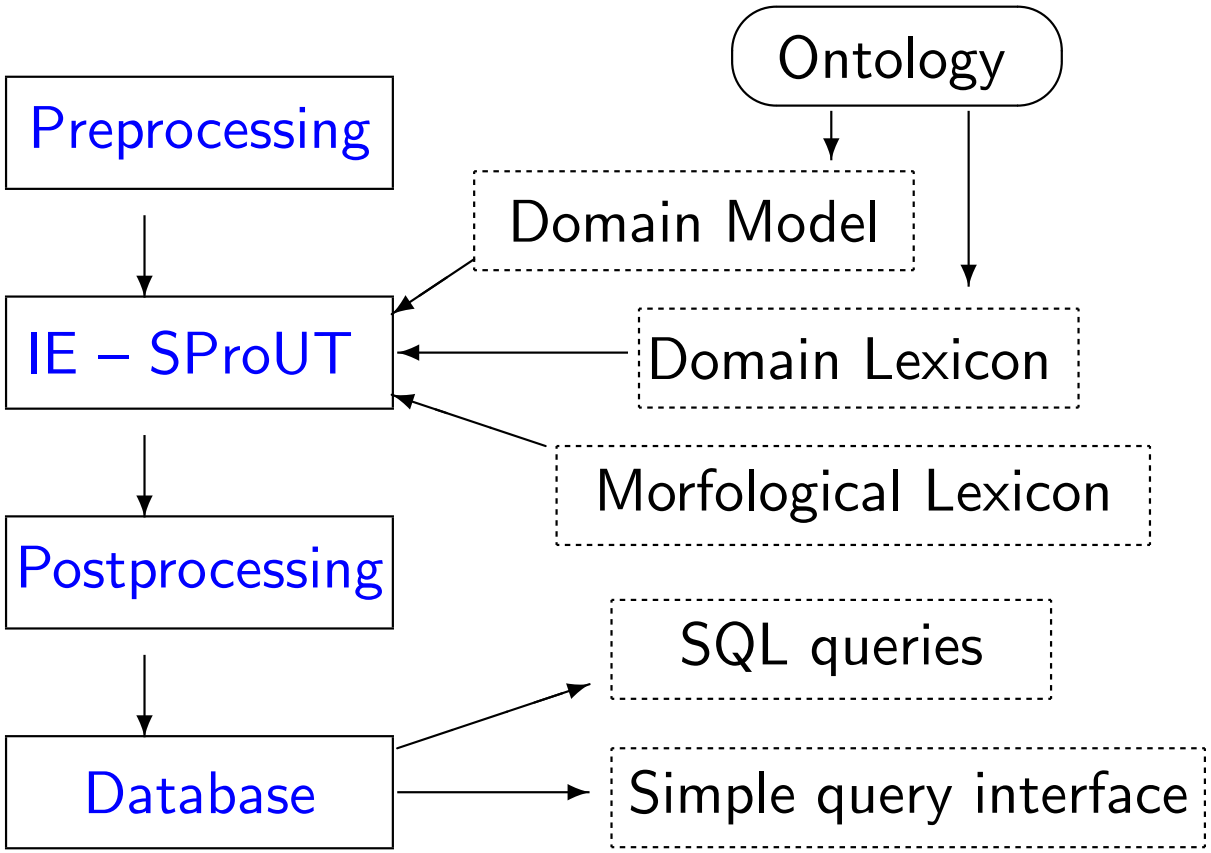
Clinical Medical Data IE System – Goals

- to test classical rule-based method on texts in Polish;
- to create relational database containing the extracted data to enable their statistical analysis or to select incomplete or irrelevant documents;
- to prepare annotated data for further IE experiments.

Two applications

1. Extracting precise information form mammography reports.
2. Selecting the most important information from diabetic patients' discharge records

System Architecture



Documents Sets Description, 1

Mammogram notes

- short texts (10-15 lines) consisting mainly of nominal phrases,
- very concise texts with many spelling and punctuation errors, nearly entire text is translated into templates slots.

Exemplary texts:

285

Sutki o resztkowym utkaniu gruczołowym z przewagą tkanki tłuszczowej. Zmian ogniskowych podejrzanych o złośliwość nie wykazano. W sutku lewym wewnątrzsutkowe węzły chłonne(prawiłowe).

286

Sutki o utkaniu z przewagą tkanki tłuszczowej. Resztkowe utkanie gruczołowe układające się pasmowato w kwadrantach górno-zewnętrznych. Zmian ogniskowych podejrzanych o złośliwość nie wykazano. Doły pachowe wolne.

Documents Sets Description, 2

Diabetes discharge reports

- 1,5 – 2,5 pages long texts in MS Word
- most data are free-form texts but some data are in tables;
- many facts are written in a standardised way, but some features can be expressed in very many ways (e.g. information about continuation of diabetes treatment);
- searched for information constitute a very small part of the input text (less then 10%).

Fragment przykładowego dokumentu

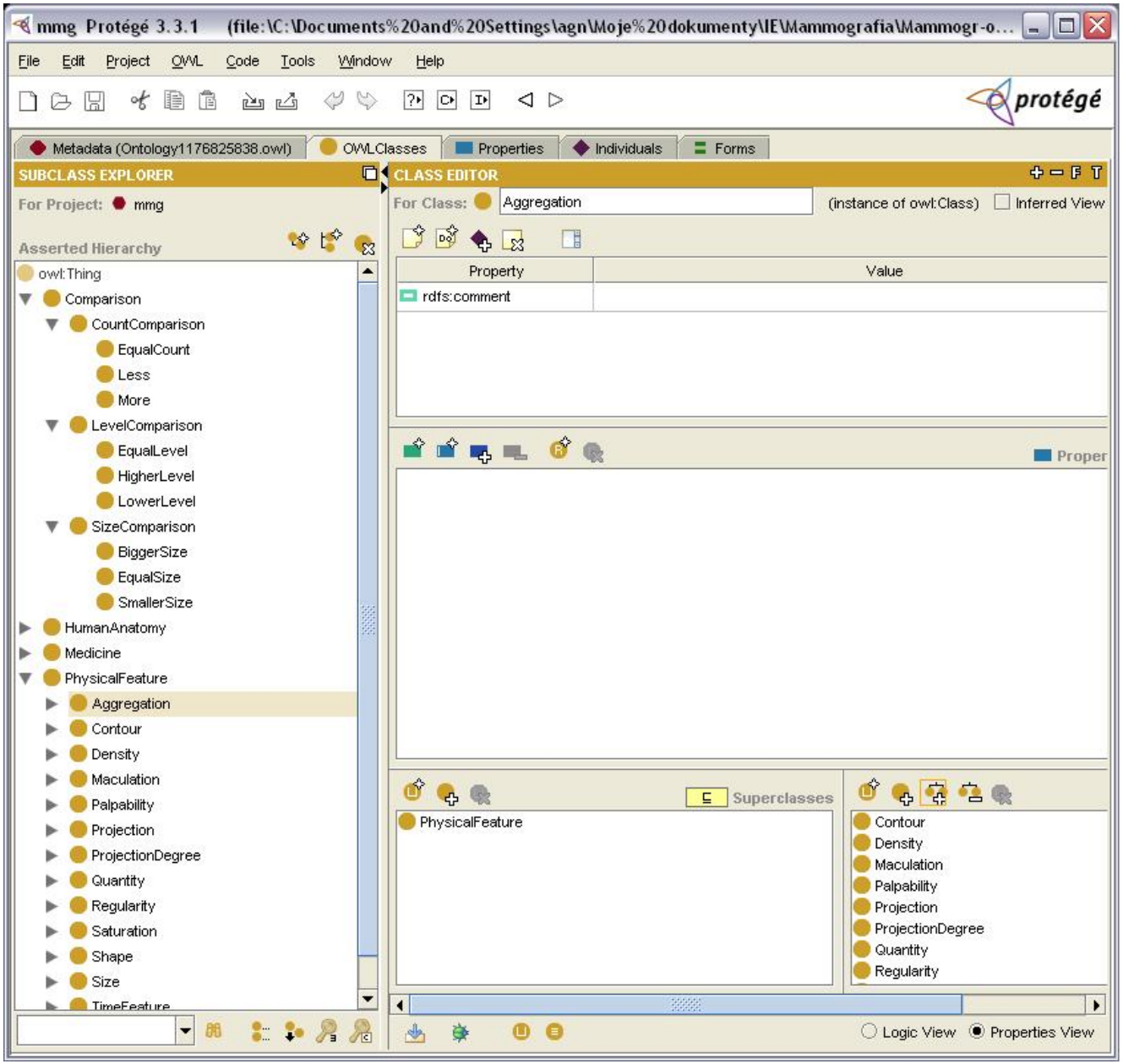
Zastosowano leczenie:
Insulina, No spa, Metocard, Prestarium, Hydroxyzyna.
Epikryza:
65 - letnia pacjentka z cukrzycą została przyjęta do kliniki z powodu znacznych wahań poziomów glukozy. Początkowo utrzymano rodzaj i system podawania insuliny razem z lekiem doustnym, który później zmieniona na insulinoterapię w trzech wstrzyknięciach na dobę. Uzyskano zadowalające poziomy glukozy w kolejnych dobowych profilach. Obserwowano wahania glikemii zarówno w ciągu dnia jak i w kolejnych dobach. Korygowano leczenie dietą- chorą przeszkolono w tym zakresie. Parametry długofalowego wyrównania są zadowalające. Pacjentka umie posługiwać się wstrzykiwaczem i glukometrem.
Nie stwierdzono późnych powikłań cukrzycy.
Kontynuowano leczenie hipotensyjne – wartości ciśnienia tętniczego były w granicach zadowalających. Konsultowana była chora przez psychiatrę z powodu utrzymującego się obniżonego nastroju. Z powodu zgłaszanych dolegliwości bólowych w nadbrzuszu wykonano badanie gastroskopowe, w którym nie stwierdzono patologii. Wydaje się że dolegliwości odczuwane przez chorą po części zależą od czynników emocjonalnych oraz czynników związanych z przebytymi operacjami (cholecystektomia i appendektomia).
Kontynuowano wcześniejsze leczenie hipotensyjne. Wartości ciśnienia tętniczego były w granicach normy.
Pacjentka zostaje wypisana do domu z zaleceniami:
1. Dieta cukrzycowa 2000 kcal., 6 posiłków/dobę. Bardzo małe drugie śniadania i podwieczorki.
2. Leki:

Domain models

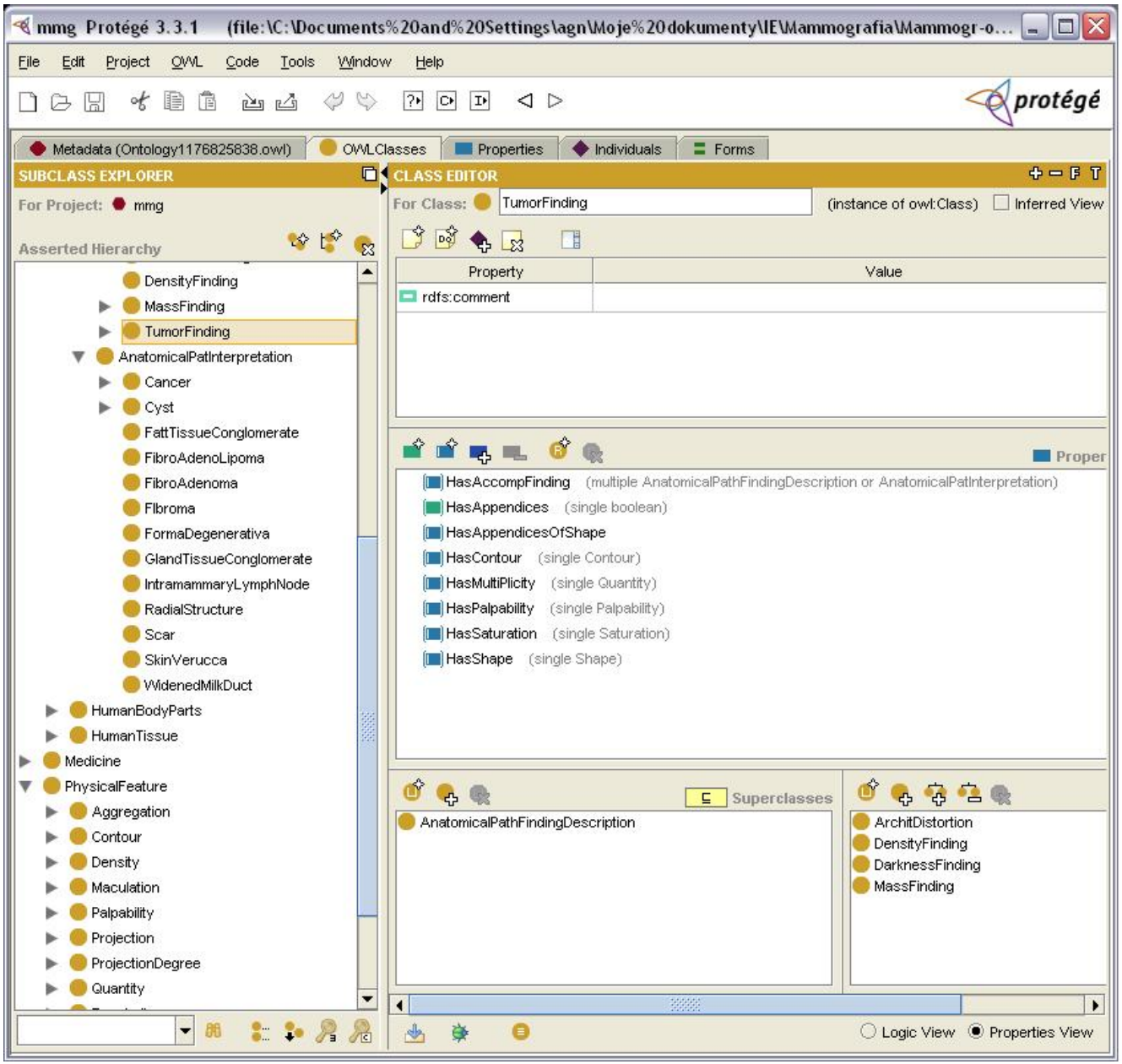
- formalized in OWL-DL standard and the Protégé ontology editor:
  - defined on the basis of an expert's knowledge and the data;
  - for diabetes ii describes information on: a patient, hospitalisation, diagnosis, complications, tests, treatment, diet;
- manually translated into TFS hierarchy (for diabetes 139 types with 65 attributes, but as much as 65 types represent medicine terms):

$$\left[\begin{array}{l} \textit{diabet\_desc\_str} \\ \text{D\_TYPE} \quad \textit{d\_type\_t} \\ \text{HBA1C} \quad \textit{string} \\ \text{UNCONTROLLED} \quad \textit{bool\_t} \\ \text{HYPOGLYCAEMIA} \quad \textit{bool\_t} \\ \text{DIAB\_FROM} \quad \textit{diab\_from\_str} \end{array}\right]$$

Ontology in Protege



Ontology in Protege



Lexicons

- Morfeusz — Morphological analyser for Polish;
- Domain Lexicon contains: names of anti-diabetic medicines, Polish and Latin domain terminology and abbreviations.

neuropatia | GTYPE: gaz_comp | G_CONCEPT: neuropathy_t
Neuropatia | GTYPE: gaz_comp | G_CONCEPT: neuropathy_t
Neuropatią | GTYPE: gaz_comp | G_CONCEPT: neuropathy_t
neuropatią | GTYPE: gaz_comp | G_CONCEPT: neuropathy_t
obwodową | GTYPE: gaz_neuro | G_CONCEPT: periphera_polineuropathy
obwodowa | GTYPE: gaz_neuro | G_CONCEPT: peripheral_polineuropathy
autonomiczną | GTYPE: gaz_neuro | G_CONCEPT: autonomic_neuropathy
autonomiczna | GTYPE: gaz_neuro | G_CONCEPT: autonomic_neuropathy

One exemplary rule

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

- *nie wykryto obecności późnych powikłań cukrzycowych pod postacią mikroangiopatii*;
- *nie występują późne powikłania cukrzycowe o charakterze mikroangiopatii*.

Negation

- *bez obecności retinopatii*
  without presence of retinopathy;
- *Nie stwierdzono późnych powikłań cukrzycy.*
  'there were no long-lasting diabetes complications';
- *Nie stwierdzono późnych powikłań cukrzycy o typie mikroangiopatii.*
  there were no long-lasting diabetes complications of microagiopathy type;
- *Nie stwierdzono późnych powikłań cukrzycy z wyjątkiem mikroangiopatii.*
  there were no long-lasting diabetes complications excluding microagiopathy;

Other Interesting issues

Coordination

- *z neuropatią autonomiczną i obwodową*
  'with autonomic and peripheral neuropathy';

Context Dependent Information

- *wystąpiła mikroalbuminuria*
  'microalbuminuria appeared' – refers to a complication;
- *Mikroalbuminuria: 25 mg/dobę*
  'Microalbuminuria: 25 mg/day' – denotes a test value

Facts expressed in many different ways

- *Wieloletnia, niekontrolowana cukrzyca typu 2*
  long-lasting uncontrolled diabetes type 2;
- *Cukrzyca wieloletnia typu 2, powikłana retinopatią i nefropatią, niekontrolowana*
  diabetes, long-lasting, type 2, recognised rethinopathy and nephropathy complications, uncontrolled.

Many Ways of Expressing the Same Information

13 different constructions recognising education:

- *Omówiono z chorym zasady diety, samokontroli i adaptacji dawek insuliny*
  'Diet, self-control and adaptation of insulin doses were discussed with the patient';
- *Nauczono chorego posługiwać sie pompą insulinową i glukometrem.*
  'The patient was taught how to use an insulin pump and a glucometer.';
- *Po odbyciu szkolenia z zakresu podstawowych wiadomości o cukrzycy wypisano chorą...*
  'After learning the basic information about diabetes, the patient was discharged...';
- *W czasie pobytu w Klinice prowadzono edukację chorej dotyczącą cukrzycy.*
  'During the patient stay in the Clinic, the patient was educated for diabetes.'....

Phrase Extraction – Evaluation of 50 Reports

	phrases	precision	recall
uncontrolled diabetes	61	100	68,85
retinopathy (total)	50	92,5	98
nonproliferative	35	100	100
preproliferative	9	100	88,89
proliferative	5	100	100
unspecified	1	20	100
diabetic education	19	100	94,74
diet modification	1	100	100

Postprocessing

Raw IE results have to be further processed. Simple postprocessing procedures consisted in:

- detection and omission of not complete or irrelevant data (reports not sufficiently filled up with data);
- omission of redundant data and choosing the most detailed information (e.g. about types of complications);
- selecting highest levels for blood test results.

More complicated procedures were needed for grouping mammography data:

- detecting the border of one pathological change description,
- delimiting partial tissue descriptions.

Mammography example

Input text

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.*)

IE results:

```
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
```

Filled template

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

AnatomicPatFindInterpret: microcalc
- Localization:
  - BodyPart:breast
  - Conventional_localization: unspec
  - Lateralization: left_right
- Number: many

RecommExam:
- ExamType: mammography
- ExamTime: year

Evaluation

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

Conclusions

- for complicated templates filling manual rule creation seems to be the proper solution,
- statistical methods can be used for improving recall.

Terminology extraction

Specialized text corpora were used also for testing methods for building domain dependent lexicons.

- construction of frequency lists for general and specialized corpora,
- building of a set of rules for recognition of 'specialized' nouns, adjectives and words unrecognized by a morphological analyzer
- importing the above defined rules into the SProUT project which recognizes noun phrases,
- using SProUT for extraction of noun phrases,
- statistical term selecting (C/NC value, Frantzi et al. 2000)

"Special" words lists

Lista słów '"specjalistycznych" tworzona jest poprzez wybór tych, dla których współczynnik względnej częstości nie przekracza 0,5:

$$r = \frac{\frac{f_{r_g}}{all_g}}{\frac{f_{r_{sp}}}{all_{sp}}}$$

f_{r_g} – liczba wystąpień słowa w korpusie ogólnym,
$f_{r_{sp}}$ – liczba wystąpień słowa w korpusie specjalistycznym,
all_g – liczba wystąpień wszystkich słów w korpusie ogólnym,
all_{sp} liczba wystąpień słów w korpusie specjalistycznym.

(dla słów występujących wyłącznie w korpusie specjalistycznym wartość współczynnika r wynosi 0)

Do dalszego przetwarzania z listy tej selekcjonowane są rzeczowniki, przymiotniki oraz słowa, które nie mają żadnego opisu w ogólnym słowniku języka polskiego (zatem nie wiadomo w szczególności do jakiej kategorii gramatycznej należą; ponad 2200 słów nie wystąpiło ani razu w korpusie ogólnym).

Candidates for terms list

Reguły rozpoznające wyróżnione "terminy":

base_n :> (morph & [STEM "zawał", SURFACE #s,
INFL infl_NOUN & [GENDER_N #g, CASE_N #c, NUMBER_N #n]] |
morph & [STEM "bolus", SURFACE #s,
INFL infl_N & [GENDER_N #g, CASE_N #c, NUMBER_N #n]] |
...)

Gramatyka rozpoznająca frazy:

n_adj-s :>
(@seek(base_n)& phrase & [AGR agr & [C #c, G #g, N #n], SURFACE #ss] |
@seek(base_tok) & phrase & [AGR agr & [C #c, G #g, N #n], SURFACE #ss])
@seek(base_adj) & phrase & [AGR agr & [C #c, G #g, N #n], SURFACE #a]
-> phrase & [AGR agr & [C #c, N #n, G #g], CAT "np", SURFACE #sa,
DTRS *cons* & [REST phrase & [SURFACE #a], FIRST phrase &
[SURFACE #ss]]], where #sa=ConcWithBlanks(#ss,#a).

Fragment of Terms List for Diabetic Reports

insulina (*insulin*)
leczenie (*treatement*)
dobowy profil glikemii (*24-hour glicaemia profile*)
ii wydział lekarski akademii medycznej
(*II Department of Public Health Care Center*)
bródnowski publiczny zakład opieki zdrowotnej
(*Bródno Public Health Care Center*)
mmol
glikemia mg
cukromocz dobowy (*24-hour urine glucose level)*

Concluding Remarks

- three different areas of the IE application have been tested,
- rule-based IE method turned out to be very efficient (good precision and recall values) for 'complicated' domains:
  - lack of annotated data,
  - great number of attributes searched for (in comparison to the amount of available texts),
  - inter connections of data and crucial dependence on negation and coordination;
- statistical techniques turned out to be good for well defined task (NER, terminology) however they also can give better results when enriched with linguistic data.

Bibliography


Kupść, A. (2006).
Extraction automatique de termes des textes polonais.
W: *Actes des 4emes Journées Internationales de Linguistique de Corpus, Lorient.*


Marciniak, M. i Mykowiecka, A. (2007).
Automatic processing of diabetic patients' hospital documentation.
W: *Proceedings of Balto-Slavonic Natural Language Processing ACL Workshop.*


Mykowiecka, A. i Marciniak, M. (2007).
Information extraction from patients' free form documentation.
W: *Proceedings of BioNLP 2007: Biological, translational, and clinical language processing ACL Workshop.*


Mykowiecka, A., Kupść, A., Marciniak, M., i Piskorski, J. (2007).
Resources for information extraction from Polish texts.
W: *Proceedings of 3rd Language Technology Conference, Poznań, 2007.*


Piskorski, J. (2005).
Named-entity recognition for Polish with SProUT.
W: *IMTCI Workshop Proceedings, Revised Selected Papers*, LNCS, str. 122. Springer-Verlag.


Piskorski, J., Sydow, M., i Kupść, A. (2007).
Lemmatization of Polish person names.
W: *Proceedings of Balto-Slavonic Natural Language Processing ACL Workshop.*