Automatic Processing of Diabetic Patients' Hospital Documentation

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

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

Abstract

The paper presents a rule-based information extraction (IE) system for Polish medical texts. We select the most important information from diabetic patients' records. Most data being processed are free-form texts, only a part is in table form. The work has three goals: to test classical IE methods on texts in Polish, to create relational database containing the extracted data, and to prepare annotated data for further IE experiments.

1 Introduction

Information extraction from natural language text has become an important task for NLP applications during the last years. In the era of huge text data collections, these methods allows us to perform searches within reasonable time. For the purpose of IE, many methods based on very different approaches (formal grammars, statistics, artificial intelligence) have already been elaborated. In spite of a great number of described experiments, the invented methods were untested on Polish texts. Nowadays, there is great interest in statistical and machine learning methods, e.g. (Bunescu et al., 2003), but applying machine learning techniques to Polish texts is difficult, as there are hardly any annotated Polish data (excluding morpho-syntactic information which is available). The second obstacle was the type of chosen data – relatively low number of available records with complex text. That is why we decided to carry out a rule-based IE system.[1] Below, we present the system selecting information from diabetic patients' hospital records written in unrestricted Polish. We defined a domain dependent set of rules basing on an expert's knowledge and tested them on the previously unseen reports. The extracted data were put into a database allowing for statistical analysis. The other result of the project, the annotated set of original reports, can be further used for applying other methods of IE to these texts.

In our project, we use the SProUT (Shallow Processing with Unification and Typed Feature Structures) system, (Drożdżyński et al., 2004). SProUT is a general purpose platform consisting of a set of components for basic linguistic operations. Grammar rules are regular expressions on typed feature structures (TFS) which are results of tokenization or morphological analysis, as well as information from the domain lexicon. SProUT differs from many other systems in that it allows for unification of TFSs thus allows more general extraction rules. Analysing Polish text is possible due to the integration (Piskorski et al., 2004) of Morfeusz, a morphological analyser for Polish (Woliński, 2006).

Although most biomedical IE activities are related to literature mining and terminology extraction, (e.g. (Bunescu et al., 2003), (Tveit and Saetre 2005)), clinical patients record mining is not a new research goal for all languages, e.g. (Hahn, Romacker and Schulz, 2002). In (Hripcsak et al., 2002) 24 clinical conditions were extracted from narrative chests radiographic reports. The task closest to the pre-

[1] Our first rule-based IE experiment concerned mammography reports (Mykowiecka, Kupść and Marciniak, 2004).

Balto-Slavonic Natural Language Processing 2007, June 29, 2007, pages 35–42,
Prague, June 2007.

sented here, i.e. searching for information contained in natural language patients' discharge summaries was undertaken in project MENELAS (Zweigenbaum, 1994) and AMBIT (Harkema et al., 2004). In the last one, the extraction rules were based on both syntactic (word category) and semantic information (e.g. latitude-noun or area-noun). 83 radiology reports were processed and descriptions of lung cancers extracted and evaluated. The exemplary results for location were: 61% precision and 92% recall. Results of our experiment are shown in sections 4 and 6.

2 Domain description

For the purpose of diabetic patients' hospital documentation analysis, we elaborated a domain model for the part of data which we are interested in. The model has been defined on the basis of an expert's knowledge and the data i.e. hospital documents. The model describes information on a patient, hospitalisation, diagnosis, tests, treatment and complications. To formalize it, we used OWL-DL standard and the Protégé ontology editor. A part of the diabetes ontology is shown in Fig. 1.

For the purpose of information extraction in SProUT, the ontology had to be translated (manually) into a typed feature structures hierarchy. In the extraction system, the domain model is represented by typed TFSs. A feature's (attribute's) value can be an atomic type, another TFS, or a list of atomic types or TFSs. The type hierarchy contains 139 types with 65 attributes, but as much as 65 types represent medicine terms.

An example of a structure defined to represent basic information about a patient's diabetes is given in Fig. 2. The structure is of the type *diabet_desc_str* and has five attributes. A value of the D_TYPE attribute has the type of *d_type_t* which is a supertype for three types of diabetes: *first, second, other*. The next attribute HBA1C refers to the results of an important diabetes monitoring test. Its numerical value is represented as a string. Next two attributes are of boolean type and indicate if the illness is uncontrolled and if the patient had incidences of hypoglycaemia. A value of the last attribute DIAB_FROM is another TFS of type *diab_from_str* representing when the diabetes have been diagnosed. This information can be given in different ways: in words e.g., *wieloletna* 'long-lasting'; as a date — *w 1990 roku* 'in the year 1990'; relatively *20 lat temu* '20 years ago'; or *w 20 roku życia* 'in the 20th year of life'. All these types of information demand different representation.

```
BiochemicalData
    BloodData
        HB1C
Diet
    DiabetDiet
DiseaseOrSymptom
    Disease
        Alcoholism
        AutoimmuneDisease
        Diabetes
            DiabetesType1
            DiabetesType2
            DiabetesTypeOther
    Symptom
        Angiopathy
            Macroangiopathy
            Microangiopathy
        BoodSymptom
            Hypoglicemia
        DiabeticFood
        Neuropathy
            AutonomicNeuropathy
            PeripheralPolineuropathy
        UrineSymptom
            Acetonuria
            Microalbuminuria
Hospitalization
Medicine
    DiabeticMedicine
        Insulin
        OralDiabeticMedicine
Treatement
    TreatementScheme
```

Figure 1: Fragment of the ontology

$$\begin{bmatrix} \textit{diabet\_desc\_str} & \\ \text{D\_TYPE} & \textit{d\_type\_t} \\ \text{HBA1C} & \textit{string} \\ \text{UNCONTROLLED} & \textit{bool\_t} \\ \text{HYPOGLYCAEMIA} & \textit{bool\_t} \\ \text{DIAB\_FROM} & \textit{diab\_from\_str} \end{bmatrix}$$

Figure 2: Structure of type *diabet_desc_str*

Every one document we process concerns one patient's visit in hospital. A particular visit is identified (see Fig. 3) by two parameters: ID number within a year and a year (attribute ID_YEAR). Sometimes some results of tests are available after the patient leaves the hospital. In such cases, there are additional hospital documents referring to these visits described by an attribute CONT: *yes* — continuation.

$$\begin{bmatrix} \textit{id\_str} & \\ \text{ID} & \textit{string} \\ \text{ID\_YEAR} & \textit{string} \\ \text{CONT} & \textit{bool\_t} \end{bmatrix}$$

Figure 3: Visit's identification structure

The specific structures are defined for representing the following information:

- identification of a patient's visit in hospital, dates when the hospitalisation took place, and its reasons,
- patient information: identification, age, sex, weight,
- data about diabetes (see Fig. 2),
- complications,
- other illnesses including autoimmunology and accompanying illnesses, which may be correlated with diabetes, like hypertension,
- diabetes treatment: recognition of insulin type and its doses and other oral medications,
- diet: how many calories, and how many meals are recommended,
- patient's education, observing of diet, modification of treatment or diet.

In order to represent complications we defined the appropriate hierarchy. It contains three main types of complications: angiopathy, neuropathy and diabetic foot. The first two have subtypes. Angiopathy divides into micro and macroangiopathy, and neuropathy can be autonomic neuropathy or peripheral polineuropathy. Micro and macroangiopaty has further subtypes. One common complication — rethinopathy is a subtype of microangiopathy and has additional attribute, that represents information about cooccurring maculopathy. Rethinopathy has also subtypes.

Sometimes it is convenient to recognise more then one complication through one rule. In this case, results are represented in a list. For example, the result of recognition of the following phrase describing complications *z neuropatią autonomiczną i obwodową* 'with autonomic and peripheral neuropathy' is represented in Fig. 4. These two complications cannot be identified separately, as there is only one occurrence of the keyword *neuropathy*.

$$\begin{bmatrix} \textit{complication\_list} & \\ \text{FIRST} & \textit{autonomic\_neuropathy} \\ \text{REST} & \begin{bmatrix} \textit{complication\_list} & \\ \text{FIRST} & \textit{peripheral\_polyneuropathy} \\ \text{REST} & \textit{null} \end{bmatrix} \end{bmatrix}$$

Figure 4: List of complications

3 Information Extraction

3.1 Domain dictionary — gazetteer

A domain dictionary contains all forms of the terms important to the domain terminology. These terms came from the data set or were introduced into the lexicon on the basis of a domain expert's knowledge. The lexicon contains among others all insulin and oral medication names important in diabetology, we introduced forms in nominative and genitive (if such exist) — only these forms appeared in the documents. The other group of words in the dictionary consists of names of diseases and diabetic complications. They have been introduced into the lexicon in all forms used in the documents.

In this specific domain lexicon, there are no information about grammatical categories because it is not used within the grammar rules. In the dictionary, we have only semantic information about entries. There are two levels of semantic information: GTYPE — groups entries with a similar meaning, and G_CONCEPT connects an entry with its unique interpretation. The lexicon is rather small — just over 200 word forms. In Fig. 5, there is a fragment of the gazetteer with eight entries. All of them refer to different types of neuropathy complications.

3.2 Grammar rules

A grammar in SProUT consists of rules, which are regular expressions over TFSs with functional operators and coreferences, representing the recognition

```
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: peripheral_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
```

Figure 5: A fragment of gazetteer

pattern. Output structures are also TFSs. Rules use three sources of information: tokenization (structures of type *token* for recognising, among others, abbreviations, dates, numbers), morphological analysis (structures of type *morph*), and a domain dictionary (*gazetteer*).

The SProUT grammar described in the paper consists of about 100 extraction rules. Each rule detects words or phrases describing information presented in section 2. For example, the rule in Fig. 6 recognises the identification number of a patient's visit in hospital. The first line recognises a word from morphological lexicon that has the base form (STEM) *numer* 'number' or an abbreviation[2] of this word, so they are recognised as a token with an appropriate SURFACE form. The next line: `token?` omits a dot after the abbreviation, if it is necessary. Next two lines recognise the keywords with the base forms *księga* ('book', 'document') and *główny* ('main'). Then, the identification number of the document is recognised by the `liczba_nat` rule called (via the `@seek` operator). The number is unified with the value of the ID attribute in the output structure. Next two lines are optional, they recognise a year number after a slash or a backslash, if this information is present. If not, the year is fixed, during postprocessing, according to the dates of the patient's visit in hospital. In this particular case, the value of the attribute CONT is *no* in the output structure. The rule in Fig. 6 captures, among other, the following phrases:

- *Numer księgi głównej 11125/2006*
  'Number of the main document 11125/2006'

[2]Abbreviations are not present in the morphological dictionary.

```
nr_ksiegi :>
      (morph & [STEM "numer"] | token & [SURFACE "nr"]
                                  | token & [SURFACE "Nr"])
      token ?
      morph & [STEM "księga"]
      morph & [STEM "główny"]
      @seek(liczba_nat) & [LICZ #nr]
      ((token & [TYPE slash] | token & [TYPE back_slash])
         @seek (liczba_nat) & [LICZ #nr1])?
->id_str & [ID #nr, ID_YEAR #nr1, CONT no].
```

Figure 6: Visit's identification rule

```
chor_autoimm:>
   (morph & [STEM "choroba"] | morph & [STEM "zespół"])
   gazetteer & [GTYPE gaz_autoimm, G_CONCEPT #type]
->autoimm_dis_str & [AUTOIMM_DISEASE #type].
```

Figure 7: Autoimmunology disease rule

- *nr księgi głównej 12354*
  'nr of the main document 12354'
- *Nr. księgi głównej 13578*
  'Nr. of the main document 13578'

The grammar rules recognising the important information are often relatively simple. There is no need to use any morphological features in rules, because we do not have to precisely delimit phrases. Searched (key) phrases consist very often of words which are very strongly connected with particular notions. For example, if we find a phrase *stopy cukrzycowej* 'diabetic foot', it is practically certain that it concerns a complication. Only base word forms (values of STEM attribute) from the morphological analyser output turned out to be necessary here.

Fig. 7 contains a simple rule recognising autoimmunology diseases. It seeks for any occurrence of the following pattern: <disease, autoimmunology-disease-specification>. The first line of the rule recognises a word: *choroba* or *zespół* e.g., 'disease'. The second line requires an entry from the domain dictionary which represents an autoimmunology disease. Its type (variable `#type`) is unified with the value of the attribute AUTOIMM_DISEASE in the output structure.

3.3 Difficult Issues

Although the results of the program are quite good, there are some difficult issues which cause errors.

We have to cope with negation, which sometimes is difficult to determine. In the following phrase: *bez obecności retinopatii* 'without presence of retinopathy', it is not enough to identify the keyword *retinopatii* 'retinopathy', it is necessary to recognise negation expressed in the form of the negative preposition *bez* 'without'. Here, the negation appeared just before the keyword, and it can be easily noticed, but sometimes a negation is far from a keyword, and is difficult to process with shallow parsing methods. Let us consider the following sentence: *Nie stwierdzono późnych powikłań cukrzycy o typie mikroangiopatii.* 'there were no long-lasting diabetes complications of microagiopathy type '. In this case, the negation *nie stwierdzono* 'there were no' is at the beginning of the sentence and the keyword *mikroangiopatii* 'microangiopathy' is the last word of the sentence. The above phrase is recognised with the rule in Fig. 8. It refers to the base forms of certain words and to the domain lexicon in order to identify a complication (variable #t). The same rule recognise, among other, the following phrases which meaning is the same as the previous one.

- *nie wykryto obecności późnych powikłań cukrzycowych pod postacią mikroangiopatii*,
- *nie występują późne powikłania cukrzycowe o charakterze mikroangiopatii*,
- *Nie stwierdzono późnych zmian cukrzycowych w postaci mikroangiopatii*.

In the very similar example: *Nie stwierdzono późnych powikłań cukrzycy z wyjątkiem mikroangiopatii.* 'there were no long-lasting diabetes complications excluding microagiopathy' the case is just the opposite, and the *microangiopathy* should be recognised. So, to properly identify whether a patient has or hasn't microangiopathy we have to analyse the whole sentence.

Some problems are caused by keywords which have different interpretation depending on the context. e.g., *mikroalbuminuria* refers to a complication in the phrase *wystąpiła mikroalbuminuria* 'microalbuminuria appeared' and denotes a test in the phrase *Mikroalbuminuria: 25 mg/dobę* 'Microalbuminuria: 25 mg/day'. In this case we determine the meaning of an ambiguous notion according to its context.

```
brak_powiklan :>
  morph & [STEM "nie"]                          ;; 'no'
  (morph & [STEM "stwierdzić"] |                ;; 'recognise'
        morph & [STEM "wystepować"] |
        morph & [STEM "wykryć"])
  (morph & [STEM "obecność"])?
  morph & [STEM "późny"]                        ;; 'long-lasting'
  (morph & [STEM "powikłanie"] |                ;; 'complication'
        morph & [STEM "zmiana"])
  (morph & [STEM "cukrzycowy"] |                ;; 'diabetes'
        morph & [STEM "cukrzyca"])
  (morph & [STEM "w"] |                         ;; preposition
        morph & [STEM "pod"] | morph & [STEM "o"])
  (morph & [STEM "postać"] |                    ;; 'type'
        morph & [STEM "typ"] | morph & [STEM "charakter"])
  gazetteer & [GTYPE gaz_comp, G_CONCEPT #t]
->no_comp_str & [N_COMP #t].
```

Figure 8: The rule recognising the lack of a specified complication

The next thing that should be taken into account, is that sometimes several pieces of information have to be recognised with one rule. In the following coordinated phrase: *retinopatię prostą oka lewego oraz proliferacyjna oka prawego z makulopatią w obu oczach* 'nonproliferative rethinopathy in the left eye and proliferative (rethinopathy) in the right eye with maculopathy in both eyes' we have to recognise both types of rethinopathy with maculopathy and create a list of complications as the output structure, see Fig. 9. The rule almost entirely refers to notions from the domain dictionary. It identifies a combination of notions denoting retinopathy. The domain dictionary contains both Polish and Latin (words in this case both languages are used by doctors) referring to this complication.

In order to recognise precisely given information, one tends to write complex rules describing entire phrases instead of separated terms. The crucial problem for the effectiveness of complex IE rules is that Polish is a free word language. This greatly increases the number of ways the same idea can be expressed. Let us consider the following examples:

- *Wieloletnia, niekontrolowana cukrzyca typu 2*, long-lasting uncontrolled diabetes type 2,
- *Niekontrolowana, wieloletnia cukrzyca typu 2*,
- *Wieloletnia cukrzyca typu 2, niekontrolowana*,
- *Cukrzyca wieloletnia typu 2, niekontrolowana*.

```
retino_koord1:>
  gazetteer & [GTYPE gaz_comp, G_CONCEPT retinopathy_t]
  token ?
  gazetteer & [GTYPE gaz_retino, G_CONCEPT #r1]
  (token){0,2}
  (token & [SURFACE “i”] | token & [SURFACE “oraz”] |
        token &[SURFACE “et”] | token & [TYPE comma])
  (gazetteer &
        [GTYPE gaz_comp, G_CONCEPT retinopathy_t])?
  token ?
  gazetteer & [GTYPE gaz_retino, G_CONCEPT #r2]
  (token){0,2}
  ((token & [SURFACE “z”] | token & [SURFACE “cum”] |
        token & [SURFACE “i”] | token & [SURFACE “oraz”])
  gazetteer & [GTYPE gaz_macul, G_CONCEPT yes & #z1 ])?
->
comp_l_str & [ COMP_L complication_list &
  [FIRST retinopathy_str &[ RETINOPATHY_T #r1 ,
           WITH_MACULOPATHY #z1 ],
  REST complication_list &
        [FIRST retinopathy_str &[ RETINOPATHY_T #r2 ,
           WITH_MACULOPATHY #z1 ],
        REST *null* ]]].
```

Figure 9: Retinopathy coordination rule

All phrases mean: ‘Long-lasting, uncontrolled, type 2 diabetes’. Every word of these phrases carries important information: *wieloletni* ‘long-lasting’, *niekontrolowany* ‘uncontrolled’, *typ 2* ‘type 2’. But they should be identified as important only in context of the keyword *cukrzyca* ‘diabetes’. The only solution is to recognise the whole phrase through one rule. So, we ought to predict all possible configurations of words and write a lot of rules that identify subsequent permutations of keywords, which might be difficult. Thus, some omissions of information can be caused by insufficient coverage by grammar rules (see sec. 4).

The information we searched for can be divided into two types. Many facts were originally written in the documents in a standardised way, for example the value of the BMI parameter, or phrases describing complications. For these parts of information, the probability of error is rather small and is related mostly to the occurrence of complicated negation or coordination. But some of the features can be expressed in many ways. In this case, the program recall can depend on the particular physicians’ writing styles. An example is the information about continuation of diabetes treatment. In this case we have to identify information about continuation of a treatment (can be expressed in many ways) in the context of a phrases denoting diabetes. This context is important because, in the texts, there are sometimes phrases describing continuation of treatment of not diabetes but other illness. A few samples are given below:

- *Kontynuowano leczenie cukrzycy dotychczasowym systemem wielokrotnych wstrzyknięć* ‘The diabetes treatment was continued on the same basis of multiple injections’,
- *Utrzymano dotychczasowy system wielokrotnych wstrzyknięć insuliny* ‘The current system of multiple insulin injections has been maintained’,
- *Kontynuowano dotychczasowy schemat leczenia cukrzycy* ‘The current schema of diabetes treatment was continued’,
- *Kontynuowano dotychczasowe leczenie hipotensyjne* ‘The current treatment of hypotension was continued’ — this phrase is not about diabetes!

A fact that a patient was educated for diabetes is another example of information which can be expressed in many ways. Any phrase indicating that a patient was informed or taught about something or something was discussed with a patient is interpreted as the information about education. We are not interested in details of education but still we have to recognise 13 different constructions describing 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.’,
- *W czasie pobytu w Klinice prowadzono edukację chorej dotyczącą cukrzycy.* ‘During the patient stay in the Clinic, the patient was educated for diabetes.’,

- *Po odbyciu szkolenia z zakresu podstawowych wiadomości o cukrzycy wypisano chorą...* 'After learning the basic information about diabetes, the patient was discharged...'.

4 IE results evaluation

Part of the data was used as a training set, the evaluation was made on the other 50 previously unseen reports. From above 60 attributes, the partial evaluation concerned only 7. The evaluated attributes are of different type: retinopathy is a keyword but we still deal with the problem of negation and coordination. Words denoting uncontrolled diabetes can refer not only to diabetes so they should be recognized only in specific contexts. Attributes: education and diet modification are represented in the texts by complex phrases.

Results are presented in Fig. 10. The worst results were observed for diabetes balance recognition. It was due to the fact that keywords representing this information had to be recognised in the context of the word *cukrzyca* 'diabetes', (see 3.3) and sometimes the distance between these words is too far. 4 occurrences of wrongly recognised *retinopathy* were caused by the unpredicted negated phrases.

	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

Figure 10: IE evaluation of 50 reports

5 Database Organization

The data obtained from the IE system is a huge XML file. The attribute values included within it were subsequently introduced into a relational database which can be searched and analysed. At the database filling stage some additional postprocessing of data was done. This concerned, among others, the following problems:

- detection and omission of information of patient not suffering from diabetes,
- detection and omission of not complete 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.

The database consists of 20 tables containing all extracted information about a patient, his/her illness and the recommended treatment. At the moment, the database contains 388 hospitalisation descriptions of 387 patients. 254 cases were qualified as diabetes type 2, 129 as type 1 and 5 as type other. 556 complications for 256 patients and 304 insulin treatment schemas have been recognised.

6 System Overview and Evaluation

The main aim of the work was creation of a system that processes diabetic patients' hospital documentation automatically and inserts the extracted data into a relational database. The database can be searched for using SQL queries or a specialized program dedicated for doctors which enables queries by example. The system architecture is given in Fig. 11. The processing procedure consisted of four stages:

- text preprocessing including format unification and data anonymization (Perl scripts),
- information extraction based on the domain model (Protégé), Polish morphological lexicon (Morfeusz) and the domain lexicon,
- postprocessing: data cleaning and structuring (Perl scripts),
- insertion data into a relational database (Postgres).

The evaluation of the system was done simultaneously with IE evaluation on the same set of 50 reports. The results are presented in Fig. 12. The final recognition of the uncontrolled diabetes was higher due to repetition of the same information in one document.

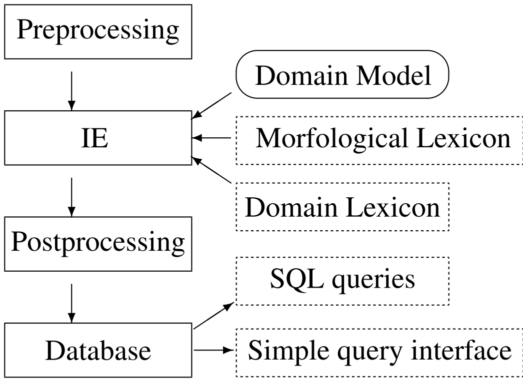

Figure 11: System architecture

	cases	precision	recall
uncontolled diabetes	37	100	86,49
retinopathy (total)	22	88	100
nonproliferative	14	100	100
preproliferative	4	100	88,89
proliferative	3	100	100
unspecified	1	25	100
diabetic education	19	100	94,74
diet modification	1	100	100

Figure 12: Overall system evaluation of 50 reports

7 Conclusions

For the chosen domain, the rule-based IE method seems to be the best one. Learning techniques are hard to apply due to: a great number of attributes searched for (in comparison to the amount of available texts) and their inter connections and crucial dependence on negation and coordination occurrences. Good precision and recall values make this method practically usable for information extraction from free patients' documentation. We plane to use our tools for creating annotated corpora (manually corrected) which are necessary for training statistical models.

Of course the portability of the method is poor. The grammars written for a particular domain can be developed to cover more facts and details but their extendibility to another domain is problematic.

Acknowledgment

This work was done in cooperation with Bródnowski Hospital in Warsaw and was partially financed by the Polish national project number 3 T11C 007 27.

References

Razvan Bunescu, Ruifang Ge, Rohit. J. Kate, Raymond J. Mooney, and Yuk Wah Wong. 2003. Learning to extract proteins and their interactions from Medline abstracts, *Proceedings of ICML-2003 Workshop on Machine Learning in Bioinformatics*, pp. 46-53, Washington, DC.

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

Udo Hahn, Martin Romacker and Stefan Schulz. 2002. Creating knowledge repositories from biomedical reports: The MEDSYNDIKATE text mining system. In *Proceedings PSB 2002*, pages 338–349.

Henk Harkema, Andrea Stzer, Rob Gaizauskas, Mark Hepple, Richard Power and Jeremy Rogers. 2005. Mining and Modelling Temporal Clinical Data. In: Proceedings of the UK e-Science All Hands Meeting 2005, Nottingham UK.

George Hripcsak, John Austin, Philip O. Alderson and Carol Friedman, 2002. Use of Natural Language Processing to Translate Clinical Information from a Database of 889,921 Chest Radiographic Reports *Radiology*.

Agnieszka Mykowiecka, Anna Kupść, Małgorzata Marciniak, 2005. Rule-based Medical Content Extraction and Classification, *Proceedings of ISMIS 2005, Springer-Verlag*.

Jakub Piskorski, Peter Homola, Małgorzata Marciniak, Agnieszka Mykowiecka, Adam Przepiórkowski and Marcin Woliński. 2004. Information Extraction for Polish using the SProUT Platform. In: *Proceedings of ISMIS 2004, Zakopane*, pp. 225–236, Springer-Verlag.

Amund Tveit and Rune Saetre, 2005. ProtChew: Automatic Extraction of Protein Names from Biomedical Literature, *Proceedings of the 21st International Conference on Data Engineering Workshops.*

Marcin Woliński. 2006. Morfeusz – a Practical Tool for the Morphological Analysis of Polish. *Procceedings of IIS: IIPWM´06*. Advances in Soft Computing, Springer-Verlag

Roman Yangarber, Winston Lin and Ralph Grishman. 2002. Unsupervised Learning of Generalized Names. *Proceedings of the 19th International Conference on Computational Linguistics, COLING 2002.*

Pierre Zweigenbaum (ed.). 1994. MENELAS: An Access System for Medical Records Using Natural Language, In: *Computer Methods and Programs in Biomedicine* vol. 45.