Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Rule-based information extraction from patients' clinical data

Agnieszka Mykowiecka *, Małgorzata Marciniak, Anna Kupść

Institute of Computer Science PAS, ul. J.K. Ordona 21, 01-237 Warszawa, Poland

ARTICLE INFO

Article history:
Received 27 May 2008
Available online 29 July 2009

Keywords:
Rule-based information extraction
Polish clinical data
Linguistic analysis

ABSTRACT

The paper describes a rule-based information extraction (IE) system developed for Polish medical texts. We present two applications designed to select data from medical documentation in Polish: mammography reports and hospital records of diabetic patients. First, we have designed a special ontology that subsequently had its concepts translated into two separate models, represented as typed feature structure (TFS) hierarchies, complying with the format required by the IE platform we adopted. Then, we used dedicated IE grammars to process documents and fill in templates provided by the models. In particular, in the grammars, we addressed such linguistic issues as: ambiguous keywords, negation, coordination or anaphoric expressions. Resolving some of these problems has been deferred to a post-processing phase where the extracted information is further grouped and structured into more complex templates. To this end, we defined special heuristic algorithms on the basis of sample data. The evaluation of the implemented procedures shows their usability for clinical data extraction tasks. For most of the evaluated templates, precision and recall well above 80% were obtained.

1. Introduction

Clinical data constitute a rich source of information about diseases, medical procedures and treatment results. It was already pointed out by several authors (e.g., [23]) that access to this information would be of a great value for clinical research and for longitudinal and epidemiological studies. However, automatic processing of clinical data is not easy.

In Polish health care institutions, examination descriptions and patient discharge records are stored in a form of text files. Although they have a common general structure, there is much variation in the way they present patient data. The content of these documents is highly dependent on the writing style of authors and varying standards. Because of that, the files are very difficult to analyze both manually (it is hard to find interesting facts in many long texts) and automatically (it is hard to specify rules for interpreting data). Thus, for administrative or epidemiological purposes, texts have to be converted into more standardized form, i.e., coded and inserted into tables or databases. An automation of this process would be very welcome because it would enhance both data accessibility and reliability (manual coding is an error-prone procedure).

Automatic processing of medical narrative data has been a research topic for decades, see [36,37]. A summary of the results achieved by 1995 is given in [41], while systems that were reported until 1999 are described in [19]. Most prominent systems that were elaborated in that period are LSP (Linguistic String Project, [36]) applied to numerous clinical domains and MedLEE [18]. The latter system is used daily at The New York Presbyterian Hospital and was adopted to process many types of medical data, among others: radiology reports, discharge summaries and mammography reports [27]. It is undoubtedly a very good example of the usefulness of MLP (Medical natural Language Processing). Apart from a few successful applications, automatic processing of clinical data is still a very challenging task. In the last ten years, it has been addressed in many projects described in various papers and books, e.g., [26,13], see bibliography in [12]. The best known projects concern processing English documents, but there is also work on processing medical documents in German [21], Dutch [42], and Bulgarian [3]. We are not aware of any similar experiment on Polish clinical data.

Natural language processing techniques can be divided into two main streams: more linguistically oriented rule-based approaches and statistical processing. Nowadays, the latter method is more popular and gained much more interest, as the results can be achieved more quickly. For example, in the biomedical domain are a great number of projects on extraction of biomedical terminology using machine learning (ML) methods (e.g., [4,47,44,5,6]). Rule-based natural language processing became less popular as it requires a lot of manual work and is not easy to reuse. Nevertheless, especially for extraction of complicated, structured templates, formal rule-based approaches are also used and give reliable results. MedLEE [13] is a rule-based system that operates on the results of a shallow syntactic analysis and refers to semantic

* Corresponding author.
E-mail addresses: agn@ipipan.waw.pl (A. Mykowiecka), mm@ipipan.waw.pl (M. Marciniak), aniak@ipipan.waw.pl (A. Kupść).

1532-0464/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2009.07.007

lexicons. Automatic learning of complicated structures is difficult especially if not much clinical data is available for training and testing, and no semantically annotated clinical data corpora are available. Nonetheless, some attempts to learn complex templates for clinical data processing are also underway. A system built by Taira and Soderland [43,40] utilizes a maximum entropy classifier for determining sentence boundaries; a lexical analyzer based on manually created lexicons with syntactic and semantic features; a statistical parser that builds dependency structures; and two kinds of semantic interpreters (rule-learning procedure and maximum entropy classifier); and finally a rule-based frame filler. To combine both symbolic and machine learning methods is probably the most promising solution. One such system is MPLUS [14] that provides syntactic analysis based on context-free grammar, and semantic analysis using Bayesian networks. One of the system's applications was for medical information extraction from Head CT reports.

Patient data can contain information on very many topics but extracting all of it would require very rich domain models. That is why in papers concerning processing and structuring clinical information usually only few selected features are addressed, e.g., [17,25,21]. In our application we extract a relatively large number of concepts – above 60. An example of work similar to ours is [11], that presents a procedure for assigning values to about 70 variables (features) by applying the MetaMap system that assigns concepts from UMLS (Unified Medical Language System) to fragments of narrative texts. That approach was later enriched by a stand alone program that can identify context changes and scope of negation [10,12]. In our extraction grammars, entire phrases expressing pertinent data are identified at once.

Similarly to others dealing with clinical data (e.g., [24]), we have decided to build a classical rule-based system. This decision was motivated by the reasons mentioned above: the complexity of templates that would be difficult to fill in using machine learning methods and the fact that no adequate annotated medical corpora are available for Polish. Nevertheless, we plan to use our IE systems to speed up the construction of annotated medical corpora that can be used for training and testing various ML methods in subsequent experiments.

IE applications for similar domains created for texts in other languages cannot be reused for Polish data. Although research on language independent applications has been carried out for some time, there are no techniques that enable the transfer of a system (especially a rule-based one) written for texts in one language to another. Additionally, in clinical applications we also have to deal with different description standards and the inflectional character of the Polish language that makes the task more complicated than for English. Therefore, for texts written in Polish we had to develop a dedicated IE system.

In the paper we present two different IE experiments.[1] The goal of our first application was to extract detailed information about breast tissue and pathological findings described in mammography reports. The second task concerned selecting crucial health information about diabetic patients from their hospital discharge documents that contain: description of state of health at the beginning and end of the hospitalization; description of patient's illnesses and their treatment; and results of all examinations.

The adopted processing strategy relies on domain knowledge and shallow linguistic analysis. The general flow of information and processing stages are shown in Fig. 1.

The organization of the rest of the paper is as follows. Section 2 includes a general description of our method. Sections 3–5 present the most interesting aspects of all processing phases, systems evaluation is provided in Section 6 while Section 7 contains a discussion of the results. Section 8 concludes the paper.

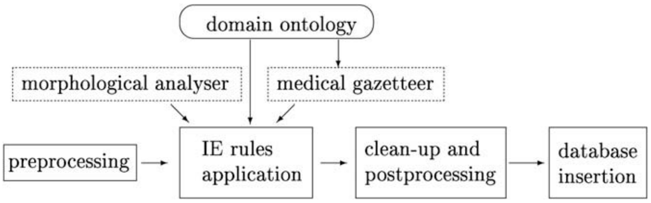

Fig. 1. Data processing stages.

2. Methods

2.1. Creating the system

The domain model and IE applications were built on the basis of an analysis of real data from Warsaw health care institutions. The mammography data consists of 2439 reports written by physicians specializing in mammogram interpretation. From this set we randomly chose a test set of 867 documents, which were not inspected during the creation of the application. The diabetic data consists of 606 hospital discharge documents from the years: 2001–2006. All documents are from a ward that specializes in treating diabetes, and are written by specialty physicians. All 169 documents from the year 2006 were not inspected until the evaluation phase.

In order to specify the information that should be extracted, we defined a medical ontology. The domain model provides a basis for structuring textual information into templates used in IE. Some ontology concepts are included in domain lexicons (gazetteers) coupled with the processing platform. The lexicons contain, among other entries, specialized terms that serve as keywords for IE. The ontology was built basing on the data and expert knowledge.

Text processing has been divided into several steps. First, the documents we obtained, required initial processing. This mainly involved spell-checking and format conversions. The core of our text analysis was done by shallow processing using a general-purpose IE platform. For processing Polish, the platform is integrated with a morphological analyzer. This allows us to deal with Polish rich inflection (7 cases, 2 numbers, 3 genders) and to use base forms of words in rules. Inflected forms of specialized terms, if they are not recognized by the morphological analyzer, are listed directly in the domain lexicons mentioned above. In IE, grammar rules aim at storing the recognized elements directly in the templates but this is not always possible. The rules take into account only a local context, whereas the required information is often scattered throughout the text. We deferred gathering all locally recognized pieces of information to the next phase, i.e., the postprocessing. At this phase, the information is completed and corrected in order to fill in more complex templates. This turned out to be a rather difficult task, especially for mammography reports, due to the complexity of the templates and the data. Finally, the templates are inserted into a relational database so that they can be easily accessed and searched.

2.2. Evaluation procedure

Both our applications were evaluated on real data. The documents used for the evaluation were randomly selected from the test data sets and consisted of 705 mammography reports and 100 diabetic discharge documents of different patients. Since manual annotation is very costly (in time and human work), we decided to apply a computer-aided annotation procedure for

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

preparing the reference set. In the first step, the test data was annotated with the corresponding IE systems, then experts corrected annotations manually (there was one expert for mammography data, and one for diabetic documents). While doing the corrections the expert looked into the original document and checked whether all information was extracted correctly. If something was missing from the results, the expert added the appropriate information to the file, and marked it with '+n'. If something was recognized incorrectly it was marked with '−n'. If an attribute was recognized properly but it was assigned the wrong value, or a block border was incorrectly placed, it was also marked. Finally, we counted all corrections made by the experts to evaluate the quality of the results of automatic extraction. We measured how well the systems recognized phrases and assigned templates. In the mammography domain, we also checked how well the block markers were inserted (not only their number but also their positions).

The evaluation procedure might be biased towards the system because the human annotator may be influenced by the automatic annotation. Despite of it, this method of verification seems to be the best when so many attributes had to be recognized – their manual recognition would take a lot of time and require an additional verification. Using this procedure, we evaluated how well our system recognizes elements of our domain model. In case of mammography reports, the evaluation of the system was also a partial and indirect evaluation of the model itself. Sometimes concepts were not recognized because we did not include them in our domain model. A few such cases occurred and all belonged to "border domains", e.g., a shadow of the pacemaker. In the diabetes domain, where our task was to recognize concepts pointed out by the potential users of the system, we were only interested in predefined attributes and their values.

For the evaluation of the presented methods we used precision, recall and their combination *F*-score. For each tested feature the measures are defined as below:

$$\text{precision} = \frac{\text{No. of phrases properly recognized as representing the feature(TP)}}{\text{No. of all phrases recognized as representing the feature(TP + FP)}}$$

$$\text{recall} = \frac{\text{No. of phrases properly recognized as representing the feature(TP)}}{\text{No. of phrases representing the feature(TP + FN)}}$$

$$F\text{-score} = \frac{2 * (\text{precision} * \text{recall})}{\text{precision} + \text{recall}}$$

Our evaluation concerned the values of the attributes and markers inserted for information grouping. In the case of attributes, we counted as TP only those attributes that were properly recognized and that were assigned proper values. Those that were assigned wrong values were counted together with wrongly recognized attributes. For computing the recall we took into account all attributes that were not recognized by our program (filled in manually by the annotator checking the output) and attributes that were assigned wrong values. In the case of markers, TP counted all markers that were inserted in the right place. FP counted both markers inserted in wrong places and markers that were completely unnecessary. FN counted all markers that were not recognized at all as well as those that were inserted in the wrong place (markers put in the wrong place were counted both as FP and FN).

3. Domain model

For any information extraction application, a proper definition of the domain model is very important. In the case of complicated medical data, this task is quite difficult and one has to decide whether to use (adapt) any of the existing models or to build a new one. Nowadays, there exist a lot of controlled vocabularies, e.g., SNOMED (http://www.ihtsdo.org/snomed-ct/) or UMLS (http://www.nlm.nih.gov/research/umls/) and ontologies e.g., BI-RADS (http://www.birads.at, [28]) that cover different areas of biomedical knowledge. Unfortunately, none of the existing (and available) resources seem to be adequate for our task. First of all they are very complicated (contain a huge number of concepts), and are not well known by Polish physicians. Learning them would require a lot of time whereas concepts used in practice by an IE application would correspond to a very small subset of these. Moreover, none of the available resources contain all concepts required for the task at hand (for a short discussion of mammography ontologies see [33]), so they would have to be modified, which is not easy for large medical ontologies and requires detailed knowledge of their structure. Another problem is that the available resources were built based on terminology different from Polish. Terminology variations are much more visible in clinical data than in scientific articles which use much more normative language. On the other hand, defining a new domain model gives full control over its scope, granularity and structure. Therefore, we decided to define our own small OWL ontology that reflects our perspective on the selected subdomain.

The created ontology [32] is based mostly on sample data and expert knowledge. An excerpt of the top levels of class hierarchy is given in Fig. 2. The picture does not show the entire topology – classes whose subclasses are not shown are marked with a black triangle. The main defined classes are *HumanAnatomy*, *Medicine* and *PhysicalFeature*. The ontology covers only subsets of the considered medical subdomains, relevant to our applications. We defined as well fragments of more general ontologies, i.e.: *PhysicalFeature*, *PhysicalFeatureComparison* and *Person*. *PhysicalFeature* class covers, among others, size, contour, aggregation, density, projection, regularity, and comparisons of quantity, degree, saturation and time. The *Time* ontology is rather simple and concerns only time expressions that occur in analyzed documents. It covers periods of time in years, months and weeks; precise and imprecise dates; and also recurrent expressions like *every year*.

Below, in (1) we present a list of properties defined for the *AnatFinding* class.

(1) *AnatFinding*:
hasAcompFinding (multiple AnatFinding)
hasAppendices (single boolean)
hasAppendicesofShape (single AppendicesShape)
hasContour (single Contour)
hasInterpretation (multiple Interpretation)
hasMultiplicity (single Quantity)
hasPalpability (single Palpability)
hasSaturation (single Saturation)
hasShape (single Shape)

The adopted IE platform, SProUT [15], cannot read OWL ontologies directly, so for both experiments we manually translated relevant parts of the ontology into a TFS (typed feature structures, [8]) hierarchy in the format required by the platform. There are two kinds of TFS: simple (atomic) and complex. A complex TFS is a typed feature structure (an attribute-value matrix) containing attributes and their values (types), see an example in (2). Each structure is assigned a type, corresponding roughly to a class from the ontology, whereas attributes correspond to properties of the class. Additionally, complex TFS structures define templates for extraction. A simple TFS consists of a type alone (there are no attributes). For example, such TFS are used to represent body parts or medication names. A complex TFS representing a template for a finding is illustrated in (2): the structure is given a type (*finding_str*) and the appropriate attributes (features). For example, the attribute ANAT_CHANGE has value of *finding_t* type whose subtypes denote different findings specified in the ontology, e.g.: tumor, density or darkness.[2]

(2)

$$\begin{bmatrix} \textit{finding\_str} & \\ \text{ANAT\_CHANGE} & \textit{finding\_t} \\ \text{CONTOUR} & \textit{contour\_t} \\ \text{MULT} & \textit{mult\_t} \\ \text{PALPABILITY} & \textit{bool\_t} \\ \text{SATURATION} & \textit{satur\_t} \\ \text{SHAPE} & \textit{shape\_t} \end{bmatrix}$$

The TFS hierarchy contains equivalents of all low-level classes (concepts) from the ontology and their properties. For example, class *Contour*, a subclass of *GeneralPhysicalFeature*, was translated into the type *contour_t*, and its subclasses were transformed into the corresponding subtypes (more specific TFS). The property *hasContour* (defined for the *AnatFinding* class) was translated into a TFS as the attribute CONTOUR that takes the value *contour_t*. The model created for the IE application is data-driven: it includes only information relevant to our applications. After the transformations, 176 types and 66 attributes for the mammography domain, and 139 types (including 75 names of medications) and 65 attributes for the diabetes domain, have been retained in the TFS hierarchy.

In SProUT, for direct binding of text strings and domain model concepts one can define specialized lexicons (gazetteers). These lexicons contain specific terms (e.g., names of medications), atypical abbreviations, and instances of key concepts, i.e., words used as triggers for IE rules. For both domains we defined such lexicons on the basis of data analysis. A small part of the gazetteer defined for the diabetes domain is given in (3). Each line corresponds to a separate lexical entry and contains the surface form (a string) and two attributes with values from the type hierarchy. The CONCEPT attribute represents the exact type of the string, while TYPE represents a more general type, usually the direct supertype in the hierarchy.

```
(3)   Neuropatia | TYPE: complication_t | CONCEPT: neuropathy_t
      neuropathie | TYPE: complication_t | CONCEPT: neuropathy_t
      obwodowa | TYPE: neuropathy_t | CONCEPT: peripheral_polyneuro
      autonomiczna | TYPE: neuropathy_t | CONCEPT: autonomic_neuro
      Gluformin | TYPE: oral_t | CONCEPT: gluformin_t
      Glurenorm | TYPE: oral_t | CONCEPT: glurenorm_t
```

As the lexicon is not coupled with a morphological analyzer, inflected forms and different spellings as they appear in the text, have to be listed directly. In order to look for a general kind of information (e.g., oral medication), that can be instantiated by different concepts in the ontology, IE rules refer to TYPE values (instead of CONCEPT). Details on how IE rules operate are presented in Section 4.

4. Extraction grammars

In our experiments, we extracted specific information using a general-purpose IE platform, SProUT, which has been adapted to process Polish, [35]. In SProUT, data at all levels of linguistic processing are represented as typed feature structures. At the lowest level, this concerns three main components:

- text tokenization (roughly, separating punctuation and splitting the text into tokens, represented as TFS structures of type *token*),
- morphological analysis (individual tokens are interpreted using a general morphological analyzer and then transformed into structures of type *morph*),
- domain lexicon lookup (each token is checked against the domain lexicon and, if a corresponding entry is found, it is associated with a TFS structure of type *gazetteer*).

Next, grammar rules operate on these entities and combine them into more complex (TFS) structures corresponding to templates. Technically, grammar rules are regular expressions over TFS, extended with unification. They are encoded using XTDL formalism, inherent to the platform. Each XTDL rule begins with a rule name. The regular expression (the rule body), placed after the ':>' symbol, describes input sequences, i.e., elements that must be identified in the text, whereas the '->' symbol indicates the resulting output structure. The regular expression can refer to the three types of input structures mentioned above: *token*, *morph* or *gazetteer*. The rule can contain an alternative ('|'), optional ('?') or repeated ('*') elements. Previously defined grammar rules can be referenced via the `@seek` operator.

Below, we present an example of the grammar rule *nr_ksiegi* recognizing an identification number of a patient's visit in hospital is given in (4). This rule captures, among others, the following phrases:

- *Numer księgi głównej 11125* 'Number of the main document 11125'
- *nr księgi głównej 12354/2006* 'No. of the main document 12354/2006'
- *Nr. księgi głównej 13578* 'No. of the main document 13578'

```
(4)  0: nr_ksiegi :>                                  ;; rule name
     1:   (token & [SURFACE "nr"] |
             token & [SURFACE "Nr"] |
     2:   morph & [STEM "numer"])                     ;; 'number'
     3:   token ?                                     ;; optional token
     4:   morph & [STEM "księga"]                     ;; 'book'
     5:   morph & [STEM "główny"]                     ;; 'main'
     6:   @seek(liczba_nat) & [LICZ #nr]              ;; number
     7:   ((token & [TYPE slash] |
             token & [TYPE back_slash])               ;; slash
     8:   @seek (liczba_nat) & [LICZ #nr1])?          ;; number
     9: ->id_str & [ID #nr, ID_YEAR #nr1].            ;; rule output
```

Lines 1–2 specify three alternative elements: a word, identified by the morphological analyzer (*morph*) via its base form (STEM) *numer* 'number', or two abbreviations of this word (identified as *tokens*). Line 3 indicates that the following token is optional: this allows us to skip the dot (a separate token) in abbreviations where it is not present (see the examples above). Lines 4–5 recognize two words via their base forms: *księga* 'book', 'document' and *główny* 'main'. Note that the alternative ('|') symbol is not inserted between the two elements and both of them have to be identified in text. Then, (line 6) the identification number of the document is recognized by a call to a different rule, `liczba_nat` (a rule that recognizes natural numbers). The recognized number is treated as a variable ('#nr') and unified with (assigned to) the value of the ID attribute in the output structure. The subsequent two lines (7–8) recognize a year ('#nr1') after a slash or a backslash, if this information is present. Finally, the output structure is provided (line 9). It contains the ID and the date (year) of the patient's visit in hospital.

[2] In all subsequent figures included in the paper names of types are omitted.

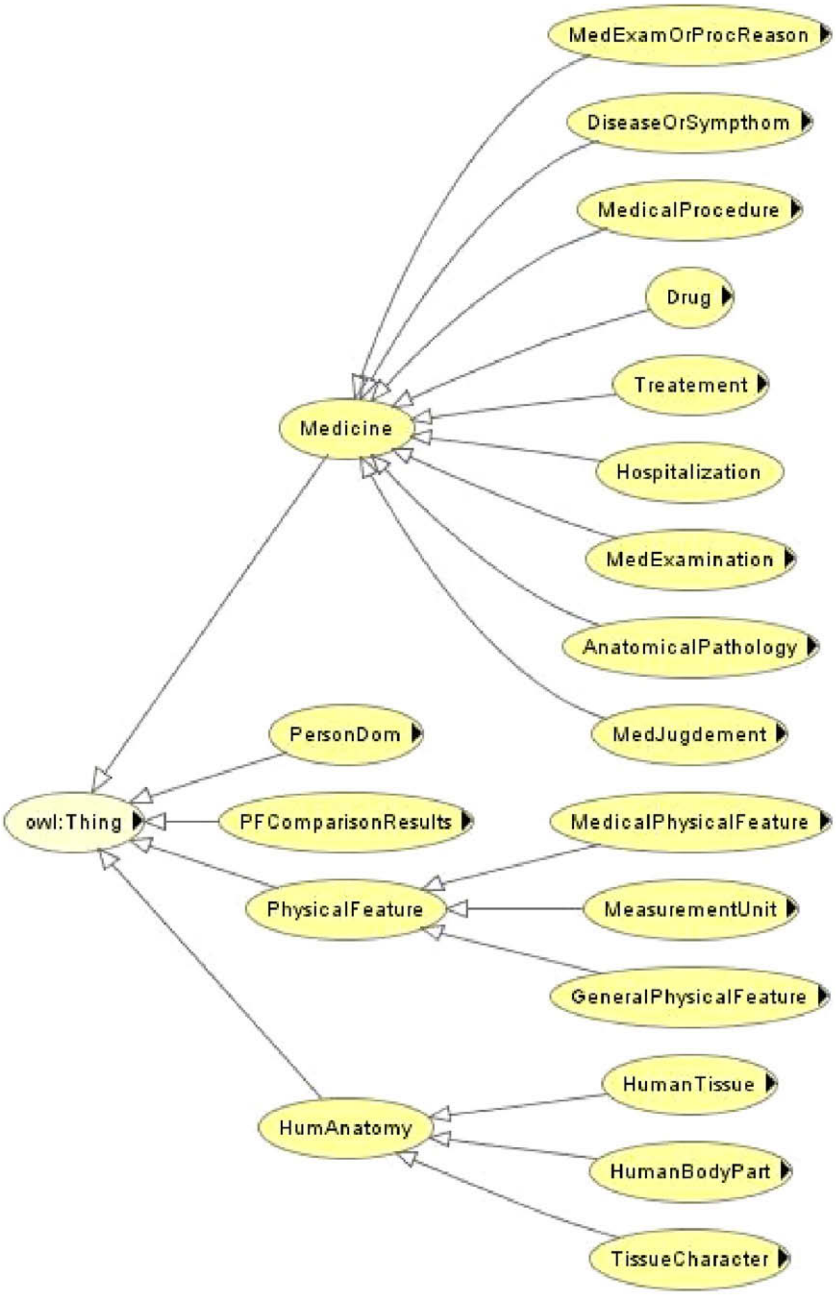

Fig. 2. Top level of the class hierarchy.

For each template specified by our domain model we define an appropriate rule (or a set of rules) that allows us to recognize its values. The grammar for mammography reports contains 190 rules, whereas that for the diabetic domain contains 150 rules. For the mammography domain nearly an entire report is covered by rules, whereas in the diabetes domain we only targeted subsets of information. Their selection was done by the physician interested in developing the application.

The extraction rules operate directly on texts but in both experiments our original data required preprocessing. For mammography reports it turned out that texts written by physicians contain a lot of spelling mistakes that decrease the quality of extracted data: a misspelled word is understandable for a human but not for an automatic processor. To address this issue, we implemented an automatic spelling correction program that uses a domain-specific lexicon [31]. Diabetic patient records were much more carefully typed, so in this case, a spell checker was not necessary and preprocessing concerned mainly format conversion (from MS Word to text files). Before using them in the experiments, all patients' records were anonymized: information about names and addresses were removed and a symbolic identification code was assigned to every patient. This process was based on the relatively strict structure of the files (personal information given in the first lines of the documents), so no sophisticated named entity recognition was applied (in contrast to [45]).

Below, we describe typical linguistic problems that had to be resolved while specifying extraction rules, and we present solutions we proposed for our applications. As rule-based extraction is strongly domain-dependent, our solutions are not general. However, many problems are common for this type of text processing and similar approaches can be adopted in many other contexts.

4.1. Ambiguous keywords

One of the main problems when interpreting natural language texts is polysemy: words are often ambiguous with respect to semantic (and morphosyntactic) interpretation. In computational linguistics the process of sense identification is called Word Sense Disambiguation (WSD). An excellent overview of WSD algorithms and applications is given in [1]. A study of WSD problems in biomedical literature and clinical notes, using a supervised machine learning approach, is presented in [39].

In our applications, the only significant difficulties were with the interpretation of keywords i.e., words important for the domain identified by the grammar rules. Some of them, e.g., insulin medication, are specified in the domain lexicon, whereas others, like *cukrzyca* 'diabetes', are referred to in the rules via their base form. While building the domain model and writing IE grammar rules, we have identified just a few such sense ambiguities. As there are no adequate tools for sense disambiguation for Polish, it had to be dealt with by our applications. On the basis of data analysis, we adopted two ways of resolving ambiguities: directly in grammar rules or in the post-processing phase.

One example of an ambiguity that can be resolved by a grammar rule is the use of the keyword *nieregularny* 'irregular': in the mammography domain, it can either describe the shape of a finding, or be interpreted as a property of the tissue. In the phrase *nieregularne zagęszczenie* 'an irregular density' the ambiguous word is found next to another keyword *zagęszczenie* 'density' that denotes a finding so the corresponding grammar rule will interpret *nieregularne* as referring to shape. An example of the same type of ambiguity in the diabetes domain is *mikroalbuminuria*, that refers to a complication in the phrase *wystpiła mikroalbuminuria* 'microalbuminuria occurred', whereas it denotes a test in the phrase *Mikroalbuminuria: 25 mg/dob*cę 'Microalbuminuria: 25 mg/day'. Therefore, when the word *mikroalbuminuria* is followed by a quantity indicating test results, the keyword is considered a test; otherwise, it is a complication.

Cases that cannot be resolved by IE grammar rules are resolved in the post-processing phase. In (5), phrases describing localization and tissue density are inserted between the keyword 'irregular' and its disambiguation context – 'glandular tissue'. In this case, the context is not taken directly into account, and the keyword 'irregular' is considered a supertype of the two possible interpretations. In the post-processing phase, if the keyword belongs to a tissue block, it is interpreted as a property of the tissue subtype, whereas if it belongs to a finding block it is meant to be a shape, cf. Section 5 for details.

(5)	[Tkanka gruczołowa]	[zabrodawkowo]	[gęsta],	[nieregularna]
	[Glandular tissue]	[subareolar]	[dense],	[irregular]

4.2. Negation

Correct recognition of negation in clinical reports is crucial to their understanding. Different approaches to the problem have been proposed. For example, [9] describes the NegEx algorithm that identifies negation with respect to findings or diseases. NegEx uses a set of regular expressions and a list of negation terms to distinguish a true negation (e.g., *absence of; declined*) from expressions that involve a negative term that does not negate the finding (e.g., *not only*). Processing negation in Bulgarian, discussed in [3], is based on a combination of shallow methods (for identification of local negation) with a deep semantic analysis (for recognition of distant negation). [16] describes dealing with negation in a system that uses the snomed-ct terminology to index clinical documents. Rules defined for the system are based on recognizing words that imply negation (e.g., *no, denies, ruled out*) and terms that stop the propagation of the assignment of negation (e.g., *other than*). In our system, recognizing both local and distant negation has been incorporated into grammar rules, occasionally increasing their complexity.

Both sets of our documents exhibit negation. If negation appears right before the keyword, it can be easily captured by a rule that recognizes a negative expression and an appropriate keyword. Unfortunately, negation does not always directly precede the keyword.

Consider the following sentence: *Nie stwierdzono późnych powikłań cukrzycy o typie mikroangiopatii.* 'there were no long-lasting diabetes complications of microagiopathy type'. The negative expression *nie stwierdzono* 'there were no' is at the beginning of the sentence, whereas the keyword *mikroangiopatii* 'microangiopathy' is the last word in the sentence. This phrase can be recognized by the rule given in (6). The rule refers to base forms of cue words and the complication (indicated by variable #t) is looked up in the domain lexicon. As a result, the complication is correctly interpreted as non-present.

```
(6)           no_complication :>
        morph & [STEM “nie”]                          ;; ‘no’
        (morph & [STEM “stwierdzić”] |
              morph & [STEM “wystepować”] |           ;; ‘recognize’
              morph & [STEM “wykryć”])
        (morph & [STEM “obecnośc̀”])?                  ;; ‘present’
        (morph & [STEM “późny”])?                     ;; ‘long-lasting’
        (morph & [STEM “powikłanie”] |                ;; ‘complication’
              [STEM “zmiana”])
        (morph & [STEM “cukrzycowy”] |
              morph & [STEM “cukrzyca”])              ;; ‘diabetes’
        (morph & [STEM “w”] |                         ;; preposition
              morph & [STEM “pod”] |
              morph & [STEM “o”])
        (morph & [STEM “postać”] |
              morph & [STEM “typ”] |                  ;; ‘type’
              morph & [STEM “charakter”])
        gazetteer & [TYPE complication_t,             ;; type of
              CONCEPT #t]                             ;; complication
     -> no_comp_str & [N_COMP #t].
```

The above rule recognizes, among others, the following phrases (with the same meaning as before):

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

In the example very similar to the last one: *Nie stwierdzono późnych powikłań cukrzycy z wyjątkiem mikroangiopatii* 'there were no long-lasting diabetes complications, except for microagiopathy', the meaning is different than in the previous examples: the 'microangiopathy' is a true complication and should be recognized. This phrase is correctly ignored by rule (6) since *z wyjątkiem* 'except for' is not a negative expression. Nevertheless, in this and similar cases, we have to analyze the whole sentence to properly identify whether a complication is or is not present, or if a certain property appears.

4.3. Coordination

In both sets of documents, there are examples of information specified by coordinated phrases. We identified three types of coordination:

- elliptical coordination,
- coordination of modifiers related to one keyword,
- coordination of keywords.

An example of elliptical coordination is the phrase *z neuropatią autonomiczną i obwodową* 'with autonomic and peripheral neuropathy', that describes two complications. The phrase consists of the keyword *neuropatią* 'neuropathy' and two coordinated adjectives indicating two types of neuropathy. Elliptical coordination has been addressed for example in [7] by machine learning techniques. As we did not have enough data for training such algorithms, we recognize such constructions within our IE grammar rules. In (7) we show the rule for recognizing coordination of neuropathies. The rule nearly exclusively refers to entries from the domain lexicon.

```
(7)   neuropat_coord:>
         gazetteer & [TYPE complication_t, CONCEPT neuropathy_t]
         (gazetteer & [TYPE neuropathy_t, CONCEPT #r1])
         (token & [SURFACE “i”]| token & [SURFACE “oraz”])
         (gazetteer & [TYPE neuropathy_t, CONCEPT #r2])
      ->feature_l_str & [FEATURE feature_list &
                            [FIRST comp_str & [COMP #r1],
                            REST feature_list & [ FIRST comp_str & [COMP#r2],
                                 REST *null* ]]].
```

The next type of coordination concerns cases when several properties are associated with one concept, e.g., many properties of a particular diabetes case, or of a finding. In (8) we show one example from the diabetes domain.

(8) *Wieloletnia powikłana retinopatią, niekontrolowana cukrzyca typu 2*
long-lasting complicated by retinopathy, uncontrolled diabetes type 2

Each word in (8) carries important information: *wieloletnia* 'long-lasting', *powikłana retinopatią* 'complicated by retinopathy', *niekontrolowany* 'uncontrolled', *typ 2* 'type 2'. Most of them are relevant to our application only in the context of the keyword *cukrzyca* 'diabetes' and should be ignored otherwise. For example, *wieloletni* in *wieloletni pacjent szpitala* 'a long-lasting patient of the hospital' does not indicate a long-lasting diabetes. On the other hand, information about retinopathy complications (also other complications or information about patient's weight) does not require any context keyword and can be correctly interpreted by itself. To convey this difference, we distinguished two groups of properties depending on the presence of the word 'diabetes' in the context. Properties that do not require this context word can be recognized separately. In this example, we recognize the whole phrase by a single rule, which allows us to fill in the template locally with up to five properties, using one keyword 'diabetes'.

The third type of coordinated phrases involves coordination of key-phrases to which the same information has to be attached. For example, in certain reports, one piece of information (e.g., a finding or diagnosis) may be associated with several localizations, see (9) and (10). The result of the analysis of (9) is shown in Fig. 3.[3] Each structure is the result of one extraction rule.

(9) *W kwadrantach górno-zewnętrznych obu sutków oraz w okolicy zabrodawkowej sutka prawego pojedyncze wyraźnie okonturowane zagęszczenia o śr 5 mm. Zmiany łagodne. (węzły chłonne? torbielki?)*
In the upper-outer quadrants and in subareolar of the right breast, there are single well circumscribed densities of a diameter of 5 mm. Benign changes (lymph nodes? cysts?)

(10) *Doły pachowe, skóra i brodawki sutkowe wolne.*
Armpits, skin and nipples are non-malignant.

In order to deal with multiple localization expressions like in (9), we introduce a grammar rule, which can handle several localizations. The rule assigns values to the LOC attribute as well as to the added LOC2 (and if necessary LOC3) attribute.

4.4. Anaphoric expressions

Anaphoric expressions, referring to an object mentioned previously in the text, have varying importance in different texts. For example, [38] suggests that in biological papers about 5% of descriptions of protein–protein interactions contain anaphoric expressions. There is much literature on anaphora and coreference resolution in texts and discourses, the Discourse Anaphora and Anaphora Resolution Conference (DAARC) is devoted to these problems. In the biomedical domain, an algorithm annotating anaphoric relations in paper abstracts is presented in [34]. The method relies on syntactic features and semantic information from UMLS. Another approach, based on supervised learning, is presented in [46], whereas a probabilistic method for anaphora resolution in full-text articles is described in [20]. Very little has been published on anaphora resolution in clinical reports. One example of a system which performs coreference resolution is MedSyndikate [22].

In our documents, interesting anaphoric expressions occurred only in mammography reports and concerned mainly the identification of localizations and a relative description of anatomic changes. Although they are not very frequent, their recognition is important for the final results. Our approach is similar to [2], as we incorporated anaphora recognition into the grammar rules. We introduce special attributes representing phrases that indicate anaphoric relations. Anaphora resolution is done at the post-processing stage. In our data, anaphoric expressions concern localization of changes or changes themselves, and can be divided into the following classes:

- anaphora of localization
    - anaphora of a general localization,
    - anaphora of a detailed localization,
- anaphora of a finding and its description
    - anaphora of a finding,
    - anaphora of finding's size,
    - anaphora of finding's features,
    - anaphoric reference to a member of a set.

An anaphora of localization is a phrase where the localization of, e.g., a finding is specified by referring to a previously mentioned

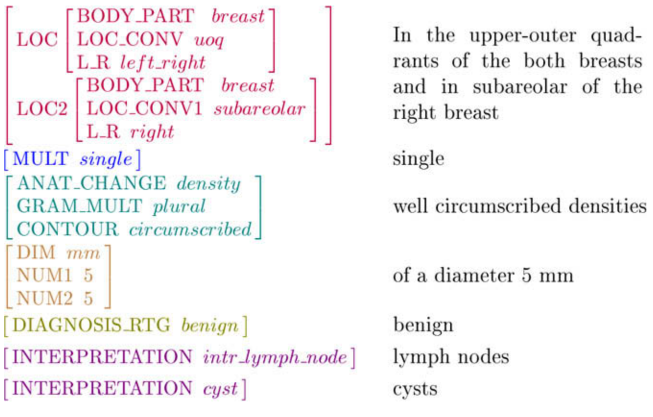

Fig. 3. Interpretation of the report cited in (9).

[3] In (9), Fig. 3, and some following figures we use colors to show to which fragment of the text each structure was assigned.

localization. The general lateralization anaphora occurs only when the lateralization information is omitted. This happens if there are two findings identified in a breast, e.g., *w kwadrancie górno-zewnętrznym tego sutka* 'in the upper-outer quadrant (uoq) of this breast', *w tym sutku w KGZ* 'in this breast in uoq'. In these phrases, the demonstrative pronouns *tym, tego* 'this$_{inst/gen}$' should be interpreted as 'in the same breast that has been described previously'. The anaphoric localization expression can also represent more detailed information. The expressions *sąsiadujący* 'neighboring', *obok* 'next to' refer to a previous localization. To represent localization anaphora, in cases where the lateralization is omitted (the first case mentioned above), the path LOC|L_R is assigned the *loc_l_r_last* type. In order to refer to the previously mentioned localization (the second case), the special LOC_REF attribute is introduced. In both cases, we calculate the appropriate localization in the post-processing phase.

The second type of anaphora concerns findings and their properties. In the following phrases: *podobna zmiana* 'a similar finding', *druga* 'the second (one)', or *zmiana o tej samej wielkości i charakterze* 'a finding of the same size and type (as the previous one)'. In such cases, the system should refer to the preceding finding in order to obtain the missing information.

In some cases only a part of a finding's description is addressed by an anaphoric expression. In (11) the second finding is mentioned by the elliptic phrase *drugie* 'the second (one)'. Localization and size of the finding are given explicitly, but other information has to be inferred from the previous finding's description.

(11) *W sutku prawym w KGW 5 mm dobrze ograniczone zagęszczenie (łagodne), drugie w KGZ o śr. 10 mm również o podobnym charakterze.*
In the right breast, in the upper inner quadrant 5 mm there is a well circumscribed density (benign), the second one in the upper-outer quadrant, of 10 mm diameter size and a similar type.

In order to account for anaphoric expressions referring to changes, we introduced three attributes, which indicate the type of referenced information. The recognition of the following phrases results in providing the corresponding structures:

- *druga zmiana* 'second finding' – [CHANGE_REF *ref_cmp*],
- *zmiana o podobnej wielkości* 'finding similar in size' – [SIZE_REF *yes*]
- *zmiana o podobnym charakterze* 'finding of a similar type' – [TYPE_REF *yes*]
- *największa z nich* 'the biggest one' – [CHANGE_REF *ref_max*].

So, the analysis of the anaphora from Example (11) gives as the result the following structure: [CHANGE_REF *ref_cmp*, TYPE_REF *yes*].

5. Post-processing – information grouping and selecting

The result of IE is an XML file that contains values of all recognized attributes in the entire document. Example (12) contains one mammography report. The result of its processing by the IE grammars is given in Fig. 4.

(12) *Sutki o resztkowym utkaniu gruczołowym w kwadrantach górno-zewnętrznych. Przewaga tkanki tłuszczowej. W obu sutkach rozsiane pojedyncze dobrze okonturowane zacienienia o śr do 11 mm. Zmiany radiologicznie łagodne (torbielki? wewnątrzsutkowe węzły chłonne?). W sutku prawym dwa makrozwapnienia o charakterze łagodnym. Doły pachowe prawidłowe. Wskazana kontrola usg.*
Breasts with the remnant glandular tissue in upper-outer quadrants. Dominant fat tissue. In both breasts spread out single well circumscribed shadows with diameter up to 11 mm. Changes radiologically benign (cysts? intramammary lymph nodes?). In the right breast there are two benign macrocalcifications. Armpits normal. USG examination recommended.

The obtained isolated pieces of information need further processing to fill in more complex templates that are defined on the basis of the domain model (ontology). For mammography reports, the most difficult task was to properly separate descriptions of different tissue types and different anatomical changes, if several findings were discovered. For example, in (12), dimension, diagnosis and interpretation structures might be attached to both finding descriptions. So, proper data grouping is essential for accurate text understanding.

Information concerning one finding is always contiguous thus grouping can be achieved by introducing beginning and ending markers. We used markers to indicate a complex template (a block) corresponding to a single finding (`fb/fe`) and a tissue structure (`tbb/tbe`).

More specific blocks in a tissue description were tagged with `tb/te`. These markers are inserted during the post-processing phase.

The post-processing phase consists of the following steps:

- identifying key attributes used for further information grouping, and markers that specify borders of three main information blocks,
- segmentation of finding block,
- dividing the description of breast composition,
- adding summarization information.

The algorithms we proposed are very strongly data dependent and were developed after analyzing many reports. The main assumption is that a description of a finding or a breast tissue can be distinguished, if we recognize the keywords indicating a le-

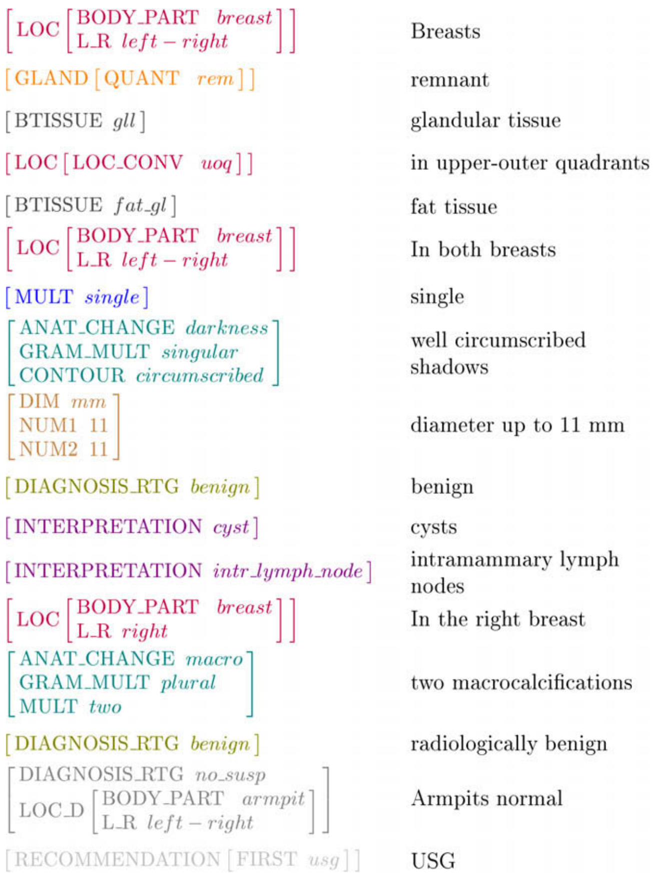
[LOC [BODY_PART breast, L_R left − right]]
Breasts
[GLAND [QUANT rem]]
remnant
[BTISSUE gll]
glandular tissue
[LOC [LOC_CONV uoq]]
in upper-outer quadrants
[BTISSUE fat_gl]
fat tissue
[LOC [BODY_PART breast, L_R left − right]]
In both breasts
[MULT single]
single
[ANAT_CHANGE darkness, GRAM_MULT singular, CONTOUR circumscribed]
well circumscribed shadows
[DIM mm, NUM1 11, NUM2 11]
diameter up to 11 mm
[DIAGNOSIS_RTG benign]
benign
[INTERPRETATION cyst]
cysts
[INTERPRETATION intr_lymph_node]
intramammary lymph nodes
[LOC [BODY_PART breast, L_R right]]
In the right breast
[ANAT_CHANGE macro, GRAM_MULT plural, MULT two]
two macrocalcifications
[DIAGNOSIS_RTG benign]
radiologically benign
[DIAGNOSIS_RTG no_susp, LOC_D [BODY_PART armpit, L_R left − right]]
Armpits normal
[RECOMMENDATION [FIRST usg]]
USG

Fig. 4. Interpretation of the mammography report cited in (12).

sion or a tissue type, respectively. The next important observation is that some features can appear in a block only once, whereas others may be repeated. We also noticed that we should start segmentation of finding blocks from the beginning of a report, while for tissue blocks a better strategy is to process data from the end. Usually, the first information about a lesion is its type (or localization) and then some details are given. In the case of tissue description the dominant tissue is usually described as last, and detailed descriptions of other tissue types are given at the beginning of a report.

In the case of information grouping key attributes for finding blocks are ANAT_CHANGE and INTERPRETATION, while BTISSUE is the key attribute for breast composition. In the preliminary phase, structures containing these attributes are additionally tagged with `a_ch`, `i_ch` and `tis` labels, respectively. Structures with attributes that cannot belong to any finding block are marked as `dloc`. Structures with information extracted from the final part of the report, containing general recommendations, are marked with the `rb` tag.

The process of identifying finding blocks starts from the first structure marked with `a_ch`, `i_ch` or `tis` tags. From that structure, we go back to the nearest block boundary or to the beginning of the report, if adding the structure to the new block is possible. Some attributes, e.g., localization or size, can appear only once in the block. Hence, if a block already contains a structure with such an attribute, it cannot be incorporated into the block. When a structure is marked with the `dloc` tag or already contains an attribute that cannot be repeated, the block beginning tag `fb` is inserted. The analogical procedure is repeated from the structure containing the key finding attribute towards the end of the report and the closing tag `fe` is inserted. Finally, the program verifies the results and corrects the segmentation if required. A pseudo-code of the algorithm (presented in [29]) is given in Fig. 5.

Results of inserting finding boundaries for the data in Fig. 4 are presented in Fig. 6.

The next post-processing step is to divide the breast's composition block, delimited by `tbb`/`tbe` tags as in Fig. 6, into logical subblocks. The algorithm is more complicated than for segmenting findings because information is not continuous.

The algorithm identifying tissue subblocks inserts `tb`/`te` tags delimiting more specific tissue information. It has been also divided into two steps. First, structures of the general composition block are annotated with tags indicating tissue type, tissue properties, or type of localization. We distinguish two tissue types: specific (glandular and glandular-fibrosis) and general (other tissue types). Additionally, we use three types of localization:

- general localization: if a LOC structure contains only the BODY_PART attribute with value *breast* and if lateralization (L_R) information is defined,

```
while not end-of-file
   // initialization
   find the beginning of the report and mark it as eb;
   copy one report to table TAB and set table TAB_TAG's rows to:
      tis if BTISSUE is found,
      a_ch if ANAT_CHANGE,
      i_ch if INTERPRETATION,
      dloc if DIAGNOSIS_RTG_LOC or attr. concerning breast composition,
   skip the last RECOMMENDATION structures and mark with rb the first one;
   skip identification information;
   checkpoint=beginning_of_report;
   // basic annotation
   while not end-of-report and not rb
      find next ut, a_ch or i_ch;
      go back to the nearest block boundary (fe, ube, dloc, eb);
      check whether unique attributes are not repeated
         and correct the boundary;
      mark the boundary as ubb or fb;
      while tag not-equal to tis, a_ch, dloc, rb
         or (i_ch if started from i_ch)
         go forward
         if unique attributes are repeated – correct the boundary;
      mark the boundary as tbe or fe;
   checkpoint=the_last_boundary+1;
   // correcting localization
   if there is a localization (LOC) outside blocks boundaries
      if the previous line is not marked i_ch and
      inside the block there is another i_ch or a_ch tag
         move the fe tag above the i_ch tag;
   if there are blocks without a localization (LOC)
      if the tbb/tbe block ends with a localization (LOC)
         move the tbe tag above the LOC line and
            redo the annotation from that line;
```

Fig. 5. Algorithm identifying findings boundaries.

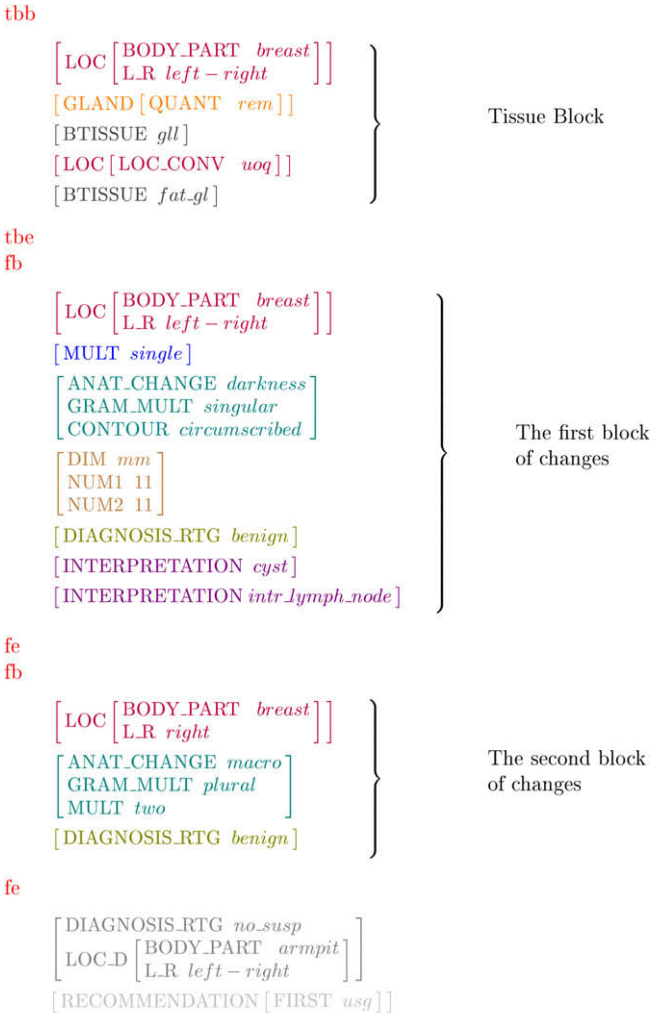

Fig. 6. Interpretation of the report with blocks.

- specific localization: if there are only attributes indicating an anatomic part or a conventional localization in the structure, i.e., LOC_CONV, LOC_CONV1 or LOC_A,
- complete localization: if both general and specific attributes are present in the structure, i.e., structure LOC has defined BODY_PART and L_R attributes as well as one of the attributes LOC_CONV, LOC_CONV1 or LOC_A.

Next, the general tissue composition block is processed from the end. Structures are grouped into logical blocks. If necessary, an appropriate localization is copied for the block describing tissue: if the block contains only specific localization, we look upward for the closest general localization; similarly, if there is a general tissue structure without any localization, we look upward for the closest general localization.

The algorithm that distinguishes specific tissue blocks, along with its problematic cases, has been described in [30]. The results obtained for the data in Fig. 6 are given in Fig. 7. The first structure, describing the general localization, is outside of any block and is not taken into account in further data processing.

In addition to attributes that are assigned values directly during text processing, three attributes were defined to provide a compact summary of the report: REPORT_CLASS – for the overall diagnosis; MMG_REL – for the image reliability; and REPORT_WITH_FIND(ings) – for a binary specification if any findings have been detected. The value of REPORT_CLASS is inferred from recommended examinations and partial diagnoses (if any) as follows:

- *diag_no_susp_changes* – no changes or other suspicious diagnoses,
- *diag_benign* – findings in the report are classified as benign,
- *diag_susp* – a biopsy is recommended or findings are classified as suspicious,
- *diag_mal* – a finding that is probably malignant is detected or an oncological consultation is required.

The reliability of the image (MMG_REL) is defined according to the type of breast composition. If the breast tissue is very dense or dysplastic, or if it is explicitly stated that the image is difficult to read, MMG_REL is *unreliable*. If the fat tissue is dominant, the report is *reliable*; in all other cases, MMG_REL takes the *avg_reliable* value.

The last stage of post-processing for mammography reports concerns transforming (copying) data for coordinated and anaphoric expressions. For example, in the following phrases: *podobna zmiana* 'a similar finding', *druga* 'the second (one)', or *zmiana o tej samej wielkości i charakterze* 'a finding of the same size and type (as the previous one)', findings and their properties are described by referential expressions. In such cases, the IE system recognizes that the finding occurs, but cannot give its characteristics. In order to account for referential expressions, we introduced three attributes: (CHANGE_REF, TYPE_REF, SIZE_REF) which indicate referential values of the selected attributes. For example, the TYPE_REF *yes* indicates that the description of the new finding is the same as that of the previous one. Similarly, SIZE_REF specifies that the size of the finding is identical to the value of SIZE or SIZE_TEXT (whichever is present) in the previous finding. In the post-processing phase, the appropriate template slots of the anaphoric finding are filled with values specified for the previous finding.

Diabetic patient's discharge documents also undergo post-processing but it is significantly simpler than for the mammography reports and concerns mainly redundant information. Below we describe the strategy for choosing important pieces of information on the same subject included in the results:

- only one copy of the information is preserved (e.g., information about uncontrolled diabetes or a complication is repeated several times in a discharge document),
- if nonidentical information on the same subject is recognized, we either:

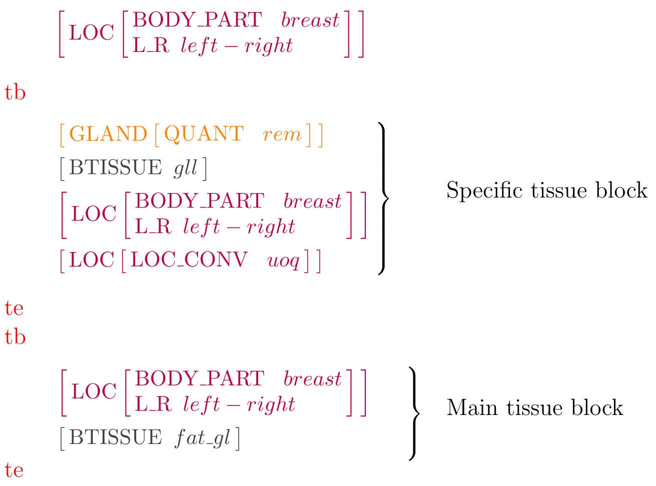

Fig. 7. Identification of breast's composition blocks.

  - preserve more specific information if this relation holds between the results, e.g., from two detected complications: *retinopathy_t* and *proliferative_retino* we choose the second one as the latter is a subtype of (is more specific than) the former,
  - choose the highest result, as for some tests with numerical values (e.g., for LDL cholesterol level),

- for medication information and recommended diet, we select information included in the summary section.

After post-processing, the filled templates are inserted into a standard relational database. This process is quite straightforward but some additional data processing is done. For example, absolute dates for relative time periods are calculated.

Table 1
Evaluation results for attributes found in 705 test mammography reports, part 1.

	Attribute	Cases	Prec.	Recall	F
ALL ATTRIBUTES		22742	99.64	99.53	99.58
Finding type	ANAT_CHANGE	277	93.81	98.56	96.13
Finding features					
Contour	CONTOUR	51	100	100	100
Saturation	SATURATION	8	100	100	100
Shape	SHAPE	39	100	100	100
Appendices shape	SH_APP	4	100	100	100
Suggestion	SUGGESTION	8	100	100	100
Accomp. calcifications	WITH_CALC	27	100	100	100
Palpability	PALPABILITY	2	100	100	100
Multiplicity					
Given by numeral	MULT	114	100	99.12	99.56
Grammatical	GRAM_MULT	287	94.10	100	96.96
Size					
Unit	DIM	103	98.10	100	99.04
Size in 1 dimension	NUM1	59	92.98	89.83	91.38
Size in 2 dimension	NUM2	67	100	100	100
Size in text	SIZE_TEXT	19	100	100	100
Visib. on other mmg pic.	PROJECTED	5	100	80.00	88.89
Interpretation	INTERPRETATION	173	98.80	95.38	97.06
Finding anaphora					
General	CHANGE_REF	13	100	100	100
Type	TYPE_REF	6	100	66.67	80.00
Size	SIZE_REF	3	100	100	100
Breast tissue type	BTISSUE	818	99.39	100	99.70
Stroma fibrosis	FIBER	17	100	100	100
Tissue features					
Density	GLAND|DENSITY	84	98.82	100	99.41
Maculation	GLAND|MACUL	71	100	98.59	99.29
Amount	GLAND|QUANT	208	99.52	100	99.76
Regularity	GLAND|REGULAR	80	100	100	100
Character	CHAR	93	98.92	98.92	98.92
Subblocks rel. comparison					
Density	CMP_DENSITY	17	100	100	100
Quantity	CMP_QUANT	7	100	85.71	92.31
Localization					
Anaphoric	LOC_REF	9	100	44.44	61.54
Body part	LOC|BODY_PART	2005	99.85	99.60	99.73
Anatomical part	LOC|LOC_A	7	83.33	71.43	76.92
Conventional part	LOC|LOC_CONV	273	100	98.90	99.45
2 dimension conv. part	LOC|LOC_CONV1	99	100	97.98	98.98
Lateralization	LOC|L_R	2009	99.85	99.55	99.70
mmg picture perspective	PROJECTION	14	100	100	100
Diagnosis	DIAGNOSIS_RTG	1517	100	99.80	99.90

Table 2
Evaluation results for attributes found in 705 test mammography reports, part 2.

	Attribute	Cases	Prec.	Recall	F
Localization with diagnosis					
Body part	LOC_D|BODY_PART	988	100	99.90	99.95
Anatomical part	LOC_D|LOC_A	61	100	100	100
Lateralization	LOC_D|L_R	989	100	100	100
Breast surgery					
Type	SURGERY	18	100	77.78	87.50
Reason	REASON	5	100	100	100
Previous exam date	PREV_EXAM	309	100	99.03	99.51
Changes from prev. exam					
In size	SIZE_CHANGE	321	100	100	100
Quantity change	QUANT_CHANGE	258	100	100	100
In saturation	SATURAT_CHANGE	321	100	100	100
Recommendations					
Examination type	RECOMMEND	769	99.35	99.87	99.61
Reason	RECOMEND_REASON	64	100	100	100
Summary features					
Report class	REPORT_CLASS	705	99.72	99.72	99.72
Where there any findings?	REPORT_WITH_FIND	705	99.57	99.57	99.57
mmg reliability	MMG_REL	705	100	100	100

6. Results

In Tables 1 and 2 we included evaluation for all attributes that occurred at least once in 705 mammography reports taken randomly from the test set.

The test set is not very big, so many attributes have a frequency equal or close to 0 (from 66 attributes 28 occurred less than 10 times). But at the same time some are quite numerous – 21 attributes occurred more then 100 times.

The results of the attribute value extraction task are very good. Twenty-eight attributes are recognized 100% correctly, 45 altogether have *F*-score above 99%. This is also the result for all attributes counted together. Only five attributes have recognition below 80%. The worst result (61.5% *F*-score for nine cases) is achieved for the recognition of relative location. For example, our grammar rules describe *w jego okolicy* ‘in its surroundings’ but they miss expressions such as *w jego obrębie* ‘within its limits’ or *wokół niego* ‘around it’.

The task of information merging is more difficult. The tissue subblocks and block beginnings are recognized with above 99% *F*-score. The worst results were obtained for separating finding descriptions (84.48% *F*-score for finding block endings) and this influences the results for tissue block ends (94.6%).

Table 3
Evaluation results for data grouping for 705 mammography reports.

	Marker	Cases	Prec.	Recall	F
Tissue block begin	tbb	706	99.86	99.58	99.72
Tissue block end	tbe	706	94.74	94.48	94.61
Tissue sub block begin	tb	966	99.69	99.90	99.79
Tissue sub block end	te	966	97.22	97.62	97.42
Finding block begin	fb	347	82.07	87.03	84.48
Finding block end	fe	347	86.38	91.35	88.80
Recommendation part begin	rb	610	99.84	99.84	99.84

Recommendations block beginnings are easy to recognize, thus the result of 99.8% *F*-score, see Table 3.

We used a manually corrected version of the system output as a reference standard. The annotations required for our task are much more detailed and complex than annotations for classification tasks, and correcting system output is one way to ensure consistency of a rater. A disadvantage of this solution is that such data are likely to present a bias towards the system. To briefly check our reference standard, we performed an experiment in which 20 mammography reports (randomly chosen but with a preference towards notes describing findings) were manually analyzed by two annotators. For these 20 reports our reference standard contains 550 attribute values. One annotator recognized 11 additional attributes' values, and omitted five, the second annotator recognized a lack of the same 11 values and omitted 16. We calculated percent agreement and Kappa between the two human annotators and between each annotator and the reference standard set on the 36 attributes that occurred at least once in the set of 20 reports, as shown in Table 4. For 20 attributes all three comparisons showed identical agreement. The Kappa coefficient for any compared pair was lower than 0.7 for four attributes. For three of them the disagreement was due to a systematic omission made by one annotator. For the fourth attribute (WITH_CALC) one annotator was better than both the second annotator and the reference standard. Only for very detailed localization description did agreement between the two annotators exceed agreement between a human and the reference annotations. Kappa coefficients between each human annotator and the reference set for this attribute were close to 0.8, whereas human agreement was equal 1. Although our reference standard set was created through a single human correcting system output, independent annotators annotating without any system feedback agreed with the reference standard set as well

Table 4
Agreement study comparing annotations from two independent annotators against each other and against reference standard annotations. A1 = Annotator 1 annotations; A2 = annotator 2 annotations; ref = reference standard annotations. PA = percent accuracy; K = Kappa coefficient.

Attribute	#ref	A1+ref		A2+ref		A1+A2	
		PA	K	PA	K	PA	K
First finding features							
ANAT_CHANGE (type)	9	0.95	0.93	0.95	0.93	1	1
CONTOUR	4	0.95	0.83	1	1	1	1
SHAPE	0	0.95	0	0.95	0	1	1
WITH_CALC	2	0.90	0.46	0.95	0.78	0.95	0.78
Multiplicity	15	0.95	0.93	0.4	0	0.40	0
Size	7	1	1	1	1	1	1
INTERPRETATION	13	1	1	1	1	1	1
Localization							
LOC|BODY_PART, LOC|L_R	16	1	1	0.95	0.94	0.95	0.94
LOC|LOC_CONV	9	0.85	0.78	0.85	0.78	1	1
LOC|LOC_CONV1	2	0.95	0.78	0.95	0.78	1	1
DIAGNOSIS_RTG	13	1	1	1	1	1	1
Second finding features							
INTERPRETATION	1	1	1	1	1	1	1
LOC|LOC_CONV	1	1	1	1	1	1	1
General breast description							
BTISSUE (tissue type)	19	1	1	0.95	0.94	0.95	0.94
LOC|BODY_PART, LOC|L_R	19	1	1	1	1	1	1
Stroma fibrosis (FIBER)	2	1	1	1	1	1	1
Glandural tissue features							
GLAND|MACUL	2	1	1	1	1	1	1
GLAND|QUANT	7	1	1	1	1	1	1
GLAND|REGULAR	4	1	1	1	1	1	1
CHAR	1	1	1	0.95	0	0.95	0
Tissue fragment features							
BTISSUE (tissue type)	9	1	1	1	1	1	1
LOC|BODY_PART,LOC|L_R	16	1	1	1	1	1	1
LOC|LOC_CONV	1	1	1	1	1	1	1
LOC|LOC_CONV1	3	1	1	1	1	1	1
PROJECTION	1	1	1	1	1	1	1
DIAGNOSIS_RTG	18	0.95	0.91	0.90	0.83	0.85	0.75
LOC_D	13	1	1	0.80	0.58	0.80	0.58
SURGERY (surgery type)	1	1	1	1	1	1	1
REASON(surgery reason)	1	1	1	1	1	1	1
PREV_EXAM(date)	11	0.95	0.94	0.80	0.74	0.85	0.81
Changes from prev. exam	10	0.95	0.92	0.95	0.92	1	1
Recommendations							
First examination	13	1	1	1	1	1	1
Time first	15	1	1	0.95	0.91	1	1
Second examination	11	1	1	1	1	1	1
Time second	3	1	1	1	1	1	1
Reason	1	1	1	0.90	0.47	0.90	0.47
ALL ATTRIBUTES	270	0.98	0.97	0.95	0.91	0.96	0.93

Table 5
Evaluation results for attributes found in 100 test discharges documents, part 1.

	Attribute	Cases	Prec.	Recall	F
ALL ATTRIBUTES		4021	99.26	96.49	97.86
Document data					
Begin	DOC_BEG	100	100	100	100
Date	DOC_DAT	100	100	100	100
Identification	ID	100	100	98	98.99
Continuation	CONT	98	100	97.96	98.97
Hospitalization from	H_FROM	99	100	98.99	99.49
Hospitalization to	H_TO	99	100	98.99	99.49
Beginning of summary	EPICRISIS	100	100	100	100
Patient data					
Identification	ID_P	100	99.01	100	99.50
Sex	ID_P_SEX	101	100	100	100
Age	ID_AGE	192	100	98.44	99.21
Height	HEIGHT	87	100	98.85	99.42
Weight	WEIGHT	88	100	100	100
BMI	BMI	88	98.82	95.45	97.11
Weight in words	W_IN_WORDS	62	98.41	100	99.20
Diabetes features					
Type	D_TYPE	173	98.84	98.27	98.55
Balance	D_CONTROL	212	98.54	95.75	97.13
Acetonuria	ACET	11	91.67	100	95.65
Treating method	D_TREAT	50	100	92	95.83
When diagnosed					
Year	Y_DAT	6	100	100	100
Relative data	D_NUM	41	100	97.56	98.77
	D_UNIT	41	100	97.56	98.77
Year of life	YEAR_OF_LIFE	3	100	100	100
In words	FROM_IN_W	32	93.94	96.88	95.38
Diseases					
All diabetic complic.	COMP	369	97.35	99.46	98.39
Retinopathy type	RETINOPATHY_T	120	98.36	100	99.17
With maculopathy	WITH_MACULOPATHY	32	100	90.63	95.08
No compl. of type	N_COMP	34	100	100	100
Accompanying	ACC_DISEASE	141	100	100	100
Autoimmune	AUTOIMM_DISEASE	4	100	100	100
Test results					
Creatinine lev. 1	CREATIN1	96	100	100	100
Creatinine lev. 2	CREATIN2	3	60	100	75
HbA1C	HBA1C	146	100	93.15	96.45
Cholesterol LDL	LDL	81	100	100	100
Microalbuminury	LEV1		100	92.59	96.15
Diet recommended					
Min. calories	CAL_MIN	102	100	94.12	96.97
Max. calories	CAL_MAX	4	100	50	66.67
Min. meals	MEALS_MIN	95	100	87.37	93.26
Max. meals	MEALS_MAX	20	100	80	88.89

Table 6
Evaluation results for attributes found in 100 test diabetes documents, part 2.

	Attribute	Cases	Prec.	Recall	F
Insulin therapy					
Dose description					
Insulin medication	I_TYPE	298	100	92.62	96.17
Min. dose	DOSE_MIN	298	100	90.60	95.07
Max. dose	DOSE_MAX	35	100	97.14	98.55
Continuous infusion					
Insulin medication	INS_TYPE	2	100	100	100
Min. basal per day	TOT_MIN_BASE	1	100	100	100
Max. basal per day	TOT_MAX_BASE	2	100	100	100
Min. bolus per meal	B_MIN	2	100	100	100
Max. bolus per meal	B_MAX	2	100	100	100
Oral medication	ORAL_TREAT	76	96.15	98.68	97.40
Various					
Reason of hospit.	REASON	83	98.73	93.98	96.30
Training of patient	EDUCATION	46	100	91.30	95.45
Beginning of insulin therapy	THERAPY_BEG	3	66.67	66.67	66.67
Therapy modification	THERAPY_ MODIF	23	100	91.30	95.45
Insulin dose modif.	DOSE_MODIF	10	90	90	90
Self monitoring	SELF_MONITORING	1	0	0	0
Diet correction	DIET_CORRECTION	1	100	100	100
Diet observing	DIET_OBSERVE	1	0	0	0

as they did with each other, indicating the quality of our reference standard set.

For the diabetes experiment, an evaluation of 55 attributes (68 total), that occurred in a test set of 100 documents is presented in Tables 5 and 6. Some attributes are rare in the test set, so interpretation of their recognition results is not reliable. Fourteen attributes occurred less than 10 times: nine of them were recognized 100% correctly; but the single occurrences of two attributes were not recognized, giving a 0% *F*-score. For the other 41 attributes: 16 have an *F*-score above 99% and three attributes have an *F*-score less than 95%. The most frequent attribute COMPlication occurred 369 times and has 98.39% *F*-score.

7. Discussion

Results of rule-based extraction systems highly depend on the quality of texts. Typographical errors, understandable for a person, for instance: 'I0 mm' instead of '10 mm', are almost impossible to fix by a computer program. Because of this, texts that will undergo automatic processing should be written very carefully. In hand-crafted grammar rules, some very typical orthographic errors, e.g., a missing space after an abbreviation, can be taken into account, but it is impossible to foresee all types of errors. For example, information about recommended diet consists of two pieces of information: how many calories and how many meals are recommended. Both values can be expressed as ranges. So finally the information is represented by four attributes. All four attributes have 100% precision but low recall mainly because of typographical problems (non-standard abbreviations, strange punctuation: six out of twelve cases of unrecognized phrases are due to lack of space between the number and the word *posiłek* 'meal'), so recall of MEALS_MIN attribute is 87.37%.

A known limitation of rule-based systems is the necessity of foreseeing all possible ways of expressing the information to be extracted. If the grammar does not cover all possibilities, the recall of the system drops, e.g., a lower recall for the INTERPRETATION attribute (94.25%) reflects the fact that some phrases used by physicians were not predicted by our grammars (e.g., abbreviation *fa* for *fibroadenoma*). This fact may also be the reason for lower precision, if some context changes in meaning are not recognized. The worst result obtained for the anatomical changes recognition (precision of 93.8%) was mainly due to the incorrect classification of previous findings as still existing (unrecognized non-local negation or previous examination contexts). Two from all three cases of unrecognized diabetes types were caused by unforeseen expressions. The phrase: *cukrzyca spowodowana leczeniem sterydami* 'diabetes caused by steroid treatment' was not recognized by our system. The second unrecognized example was *cukrzyca typu 3* 'diabetes type 3'.

The lower precision obtained for the task of identifying finding block beginning was partially due to errors in recognition of ANAT_CHANGE and INTERPRETATION features and some localization recognition errors. The other important reason of incorrect finding description borders insertion was incorrect separation of a tissue description from a finding description segment. This error is relatively easy to make, as very often the tissue changes into a description of a finding very smoothly without repetition of the localization that crosses borders.

A rule-based system is not able to recognize a new name until it is explicitly defined. For example, 3 new names of insulin medications results in 9 unrecognized cases of I_TYPE attribute. In the case of IE systems based on statistical methods, it might be possible to recognize unknown types of insulin, simply because the system could identify them on the basis of available examples through training. Statistical methods are good at recognizing simple information such as diseases occurrence, but it is not straightforward to indicate which of them are diabetic complications. Rule-based methods are better at recognition of rare and complex information, for example description of continuous insulin infusion therapy (14 examples in training data), which includes: insulin type; description of basal dose described as a dose per hour or a total daily dose; and description of a bolus dose per meal or doses for particular meals. Both approaches seem to be complementary to a certain degree so, probably the best solution is a combination of a rule-based approach with some statistical methods.

Note that the procedure of developing the reference set can influence results. As we have already mentioned a set based on automatically annotated data is biased towards the system – a human corrector tends to preserve system decisions and may not notice facts which were omitted by the system. On the other hand, manual annotation can be inconsistent in the way of annotating the same information and in the decisions of what is and what is not worth annotating. The more detailed the annotation guidelines, the closer manual annotation is to the automatic annotation. To evaluate our reference standard we performed a selective manual annotation study, which showed that in this particular case the system-aided annotation was very close to that achieved manually.

8. Conclusions

The goal of the presented research was to determine whether IE techniques based on knowledge about a particular natural language and application domain combined with domain specific post-processing procedures, can give satisfactory results in the extraction of data from free-text clinical documents. In the paper we described an IE system that encodes detailed information from medical texts written in Polish. Even for English texts there are only a few MLP systems that offer this amount of detail.

Because of the complexity of templates and lack of training data we chose a rule-based IE method. This method, although already used in many applications, has not yet been applied to Polish data. It was interesting to see how the diversity of word forms and a relatively free word order can influence the usability of IE techniques. In our system, the variability of inflectional forms had to be addressed through the use of both: a general lexicon and a domain lexicon that include various morphological word forms (even medication names can be inflected). Free word order had to be accounted for in grammar rules (several permutations are necessary) or in the post-processing stage (when different surface realizations disallowed information merging directly in the grammar).

Creating a rule-based system requires a lot of domain knowledge and is time consuming, but we are convinced that the results shown in the paper support the thesis that such systems can be reliable and useful for automated clinical data processing. After thorough testing and adapting to particular data sources they can be applied to historical documents to extract data for statistical purposes. In particular, they allow for the extraction of information that is included in these files, but that was previously treated as less important and was not put into structured databases. The presented method can also be used on currently gathered patients' data to extract the most important facts, and to select groups of patients that need special attention. Additionally, a higher documentation quality can be achieved by using such systems to check the consistency of tabular and text data, and to look for documents in which a specific type of information is missing.

References

[1] Agirre E, Edmonds P, editors. Word sense disambiguation. Algorithms and applications. Text, speech and language technology. Springer-Verlag; 2007.

[2] Bontcheva K, Dimitrov M, Maynard D, Tablan V, Cunningham H. Shallow methods for named entity coreference resolution. In: Proceedings of traitement automatique des langues naturelles. TALN Nancy; 2002.

[3] Boytcheva S, Strupchanska A, Paskaleva E, Tcharaktchiev D. Some aspects of negation processing in electronic health records. In: Proceedings of international workshop language and speech infrastructure for information access in the Balkan countries; 2005. p. 1–8.

[4] Bunescu R, Ge R, Kate RJ, Mooney RJ, Wong YW. Learning to extract proteins and their interactions from medline abstracts. In: Proceedings of ICML-2003 workshop on machine learning in bioinformatics; 2003. p. 46–53.

[5] Burnside E, Rubin D, Shachter R. A Bayesian network for mammography. In: Proceedings of the American medical informatics association symposium; 2000. p. 16–110.

[6] Burnside E, Strasberg H, Rabin D. Automated indexing of mammography using linear least squares fit. In: CARS 2000 international conference on computer assisted radiology and surgery, San Francisco, CA; 2000.

[7] Buyko E, Tomanek K, Hahn U. Resolution of coordination ellipses in biological named entities using conditional random field. In: Proceedings of the 10th conference of the Pacific association for computational linguistics; 2007.

[8] Carpenter B. The logic of typed feature structures. Cambridge tracts in theoretical computer science. Cambridge University Press; 1992.

[9] Chapman WW, Hanbury B, Cooper P, Buchanan G. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform 2001;34:204–30.

[10] Chapman WW, Dowling JN, Chu D. Context: an algorithm for identifying contextual features from clinical text. In: ACL 2007. Proceedings of the workshop on BioNLP 2007; 2007.

[11] Chapman WW, Fiszman M, Dowling JN, Chapman BE, Rindflesch TC. Identifying respiratory findings in emergency department reports for biosurveillance using metamap. In: Medinfo; 2004. p. 487–91.

[12] Chapman WW, Dowling JN, Ivanov O, Gesteland P, Olszewski R, Espino J, et al. Evaluating natural language processing applications applied to outbreak and disease surveillance. In: Proceedings of 36th symposium on the interface: computing science and statistics, Baltimore, MD; 2004.

[13] Chen H, Fuller S, Friedman C, Hersh W, editors. Medical informatics: knowledge management and data mining in biomedicine. Springer; 2005.

[14] Christensen LM, Haug PJ, Fiszman M. MPLUS: a probabilistic medical language understanding. In: Proceedings of ACL workshop on natural language processing in the biomedical domain; 2002.

[15] Drożdżyñski W, Krieger H-U, Piskorski J, Schäfer U, Xu F. Shallow processing with unification and typed feature structures – foundations and applications. German AI J KI-Zeitschrift 2004;01/04.

[16] Elkin PL, Brown SH, Bauer BA, Husser CS, Carruth W, Bergstrom LR, et al. A controlled trial of automated classification of negation from clinical notes. BMC Med Inform Decision Making 2005;5.

[17] Fiszman M, Chapman W, Aronsky D, Evans S, Haug P. Automatic detection of acute bacterial pneumonia from chest X-ray reports. J Am Med Inform Assoc 2000;7.

[18] Friedman C, Alderson P, Austin J, Cimino J, Johnson S. A general natural language text processor for clinical radiology. J Am Med Inform Assoc 1994;1.

[19] Friedman C, Hripcsak G. Natural language processing and its future in medicine. Acad Med 1999;74(8):890–5.

[20] Gasperin C, Briscoe T. Statistical anaphora resolution in biomedical texts. In: Proceedings of the 22nd international conference on computational linguistics (Coling 2008), Manchester, UK, August 2008. Coling 2008 Organizing Committee; 2008. p. 257–64.

[21] Hahn U, Romacker M, Schultz S. MEDSYNDIKATE – a natural language system for the extraction of medical information from findings reports. Int J Med Inform 2002;67:63–74.

[22] Hahn U, Romacker M, Schulz S. Creating knowledge repositories from biomedical reports: the MEDSYNDIKATE text mining system. Pac Symp Biocomput 2002;338–49.

[23] Harkema H, Gaizauskas R, Hepple M, Roberts A, Roberts I, Davis N, et al. A large scale terminology resource for biomedical text processing. In: HLT-NAACL 2004 workshop: BioLINK 2004 linking biological literature. Ontologies and databases; 2004.

[24] Harkema H, Roberts I, Gaizauskas R, Hepple M. Information extraction from clinical records. In: Proceedings of the 4th UK e-Science all hands meeting, Nottingham, UK; 2005.

[25] Hripcsak G, Austin J, Anderson P, Friedman C. Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports. Radiology 2002;224:157–63.

[26] Hripcsak G, Friedman C, Alderson PO, DuMouchel W, Johnson SB, Clayton PD. Unlocking clinical data from narrative reports: a study of natural language processing. Ann Intern Med 1995;122(9):681–8.

[27] Jain NL, Friedman C. Identification of findings suspicious for breast cancer based on natural language processing of mammogram reports. In: Proceedings of the American medical informatics association annual fall symposium; 1997. p. 829–33.

[28] Kopans DB, D'Orsi C, Adler D, Bassett LW, Brenner RJ, Dodd GD, et al. Breast imaging reporting and data system (BI-RADS). In: American College of Radiology, Reston, Va; 1993.

[29] Marciniak M, Mykowiecka A, Kupść A, Piskorski J. Intelligent content extraction from Polish medical reports. In: International workshop on intelligent media technology for communicative intelligence. LNCS 3490. Springer-Verlag; 2005.

[30] Mykowiecka A, Kupść A, Marciniak M. Rule-based medical content extraction and classification. In: Proceedings of ISMIS 2005, Gdańsk. Springer-Verlag; 2005.

[31] Mykowiecka A, Marciniak M. Domain-driven automatic spelling correction for mammography reports. In: Proceedings of intelligent information systems 2006. p. Springer.

[32] Mykowiecka A, Marciniak M. Domain model for medical information extraction – LightMedOnt ontology. In: Aspects of natural language processing, Festschrift in Honor of L. Bolc, LNAI 5070. Springer-Verlag. p. 341–67.

[33] Mykowiecka A, Marciniak M, Podsiadły-Marczynkowska T. A data-driven ontology for an information extraction system from mammography reports. In: Proceedings of 10th International Protégé conference, Budapest; 2007.

[34] No JC, Zhang J, Pustejovsky J. Anaphora resolution in biomedical literature. In: Proceedings of international symposium on reference resolution for NLP, Alicante, Spain; 2002.

[35] Piskorski J, Homola P, Marciniak M, Mykowiecka A, Przepiórkowski A, Woliński M. Information extraction for Polish using the SProUT platform. In: Intelligent information processing and web mining. Proceedings of the IIS: IIPWM'04. Springer-Verlag; 2004.

[36] Sager N, Friedman C, Lyman M. Medical language processing: computer management of narrative data. Addison-Wesley Publishing Company; 1987.

[37] Sager N, Lyman M, Bucknall C, Nhan N, Tick LJ. Natural language processing and the representation of clinical data. J Am Med Inform Assoc 1994;1(2):142–60.

[38] Sanchez-Graillet O, Poesio M, Kabadjov M, Tesar R. What kind of problems do protein interactions raise for anaphora resolution? – a preliminary analysis. In: Proceedings of the second international symposium on semantic mining in biomedicine, Jena, April 9–12, 2006.

[39] Savova G, Coden A, Sominsky I, Johnson R, Ogren P, Groen P, et al. Word sense disambiguation across two domains: biomedical literature and clinical notes. J Biomed Inform 2008;41(6):1088–100.

[40] Soderland SG. Building a machine learning based text understanding system. In: Proceedings of IJCAI-2001 workshop on adaptive text extraction and mining; 2001.

[41] Spyns P. Natural language processing in medicine: an overview. Methods Inform Med 1996;35:285–301.

[42] Spyns P, Nhan TN, Baert E, Sager N, Moor GD. Medical language processing applied to extract clinical information from Dutch medical documents. In: Proceedings of the 9th world congress on medical informatics (MEDINFO98); 1998. p. 685–9.

[43] Taira RK, Soderland SG, Jakobovits RM. Automatic structuring of radiology free-text reports. Radiographics 2001;21:237–45.

[44] Tveit A, Saetre R. Protchew: automatic extraction of protein names from biomedical literature. In: Proceedings of the 21st international conference on data engineering workshops; 2005.

[45] Uzuner O, Sibanda TC, Luo Y, Szolovits P. A de-identifier for medical discharge summaries. Artif Intell Med 2007;42:13–35.

[46] Yang X, Su J, Zhou G, Tan CL. An np-cluster based approach to coreference resolution. In: Proceedings of Coling 2004, Geneva, Switzerland, August 23–27, 2004; COLING. p. 226–32.

[47] Yangarber R. Acquisition of domain knowledge. In: Pazienza T. (editor). Information extraction: in the Web Era. Natural language communication for knowledge acquisition and intelligent information agents, LNAI 2700. Springer-Verlag; 2003. p. 1–28.