Structuring of Status Descriptions in Hospital Patient Records

Svetla Boytcheva\textsuperscript{1}, Ivelina Nikolova\textsuperscript{2}, Elena Paskaleva\textsuperscript{2}, Galia Angelova\textsuperscript{2}, Dimitar Tcharaktchiev\textsuperscript{3} and Nadya Dimitrova\textsuperscript{4}

\textsuperscript{1}State University of Library Studies and Information Technologies, Sofia, Bulgaria
\textsuperscript{2}Institute for Parallel Processing, Bulgarian Academy of Sciences, Sofia, Bulgaria
\textsuperscript{3}USHATE, Medical University, Sofia, Bulgaria
\textsuperscript{4}National Oncological Hospital, Sofia, Bulgaria

May 18, 2010
Overview

1. IE in the EVTIMA System
2. System Functionality
3. Evaluation
Motivation

Why do we process automatically Personal Records?

- Fast growing archives of medical information which could be a source for finding dependencies observable only on a big scale;
- The extracted and structured features will be employed for classification of patient cases, effective information retrieval and further processing in different medical tasks;
- EVTIMA stands for "Effective search of conceptual patterns with applications in Medical Informatics"
- IN A LONG RUN: For gaining better quality of health care, health management etc.
Resources

Available resources

- A lexicon of 30,000 Bulgarian lexemes
- The International Classification of Diseases (ICD-10) (10,970 terms) without clinical modification and ICD-9 with CM
- Official list of the registered in Bulgaria drugs maintained by the Bulgarian Drug Agency (5,600 drug names in Latin)

Resources under development

- Syntactic rules for partial analysis
- List of drugs registered in Bulgaria (in Bulgarian)
- Semantic network encoding concepts (organs, diseases, symptoms, treatments) and conceptual relations among them with ontological vocabulary in Bulgarian
- … actually we report work in progress
The medical Language I

PRs’ status texts characteristics:
- PRs are 2-3 pages long, split in 8 zones (some of them may be omitted);
- Contain a specific mixture of terminology in Latin (about 1%) and Cyrillic letters;
- Contain Latin terms transcribed with Cyrillic letters;
- Terms occur in the text with a variety of word forms - due to the highly-inflectional Bulgarian language;
- Consist of short declarative sentences and sentence phrases without agreement, often without proper punctuation marks;
PRs’ status texts characteristics:

- Descriptions are often missing:
  - 86% of the PRs in our corpus discuss explicitly the patient status regarding skin colour,
  - 63% - fat tissue,
  - 42% - skin turgor and elasticity,
  - 13% - skin hydration;

- Contain numerical values of analyses and clinical test data (about 16%);

- Contain various typos.
1. Selecting anatomic organ for processing
IE Steps

1. Selecting anatomic organ for processing

2. Finding relevant characteristics, features and related organs
1. Selecting anatomic organ for processing

2. Finding relevant characteristics, features and related organs

3. Determining the scope of the expression for the chosen anatomic organ
IE Steps

1. Selecting anatomic organ for processing

2. Finding relevant characteristics, features and related organs

3. Determining the scope of the expression for the chosen anatomic organ

4. Finding the most greedy regular expression
IE Steps

1. Selecting anatomic organ for processing

2. Finding relevant characteristics, features and related organs

3. Determining the scope of the expression for the chosen anatomic organ

4. Finding the most greedy regular expression

5. Applying additional rules for analysis
IE Steps

1. Selecting anatomic organ for processing

2. Finding relevant characteristics, features and related organs

3. Determining the scope of the expression for the chosen anatomic organ

4. Finding the most greedy regular expression

5. Applying additional rules for analysis

6. Using regularity rules about attributes correlations
IE Steps

1. Selecting anatomic organ for processing

2. Finding relevant characteristics, features and related organs

3. Determining the scope of the expression for the chosen anatomic organ

4. Finding the most greedy regular expression

5. Applying additional rules for analysis

6. Using regularity rules about attributes correlations

7. Choosing template slots to be filled in
IE Steps

1. Selecting anatomic organ for processing

2. Finding relevant characteristics, features and related organs

3. Determining the scope of the expression for the chosen anatomic organ

4. Finding the most greedy regular expression

5. Applying additional rules for analysis

6. Using regularity rules about attributes correlations

7. Choosing template slots to be filled in

8. Filling in default values in the obligatory template fields which remained empty
IE Steps 1

Selecting a feature for processing - thyroid gland, skin, limbs/legs, eyes, neck etc.:

1. Zones segmentation;

2. Mining the anamnesis - age, sex, sickness name, type and duration;

3. Mining the patient status zones - limbs, skin and thyroid gland condition;

4. Choosing relevant feature slots for the selected organ;

5. Identifying the description boundaries for each particular organ by using domain knowledge.
Example

Крайници - отслабени пулсации на a. dorsalis pedis двустранно. Претибиални и перималеоларни отоци. Онихомикоза, tinea pedis. Сукусио реналис - (-) отр. двустранно.

*Lims - reduced dorsal pedal pulse on both feet. Pretibial and perimaleolar oedema. Onychomycosis, tinea pedis. Succusio renalis - bilateral negative (-).*
Semantic network

- artery
- peripheral_artery
- right_dorsalis_pedis
- lower_right_limb
- lower_left_limb
- lower_lim
- upper_right_limb
- upper_left_limb
- left_leg
- left_shank
- left_foot
- left_hand
- finger
- nail
- has_location
- disease
- infection
- onychomycosis
- tinea_pedis
- part_of
- isa
- has_location
Example

Limbs - without oedema, varicose changes. Preserved peripheral arteries pulse, preserved tactile, thermo and vibratory sensation. Walks with difficulty, algetic gait, uses assistive devices.
IE Steps

1. Selecting anatomic organ for processing
   ↓
2. Finding relevant characteristics, features and related organs
   ↓
3. Determining the scope of the expression for the chosen anatomic organ
   ↓
4. Finding the most greedy regular expression
   ↓
5. Applying additional rules for analysis
   ↓
6. Using regularity rules about attributes correlations
   ↓
7. Choosing template slots to be filled in
   ↓
8. Filling in default values in the obligatory template fields which remained empty
Regular Expressions

- Description of one AO, all its characteristics and their features presented in one sentence:

  AO [-] ['with'/'of' F] Ch1, ['with'/'of' F] Ch2, ...

  "Крайници без отоци, запазени периферни пулсации, онихомикоза." (Lower limbs without oedema, preserved peripheral pulse, onychomycosis)

- Description of one AO, all its characteristics and their features presented in several consecutive sentences:

  AO [-] ['with'/'of' F] Ch1. ['with'/'of' F] Ch2. ...

  "Крайници - без отоци. Запазени пулсации на периферните артерии." (Lower limbs without oedema. Palpable peripheral arteries pulse).

  etc.
IE Steps

1. Selecting anatomic organ for processing
   ↓
2. Finding relevant characteristics, features and related organs
   ↓
3. Determining the scope of the expression for the chosen anatomic organ
   ↓
4. Finding the most greedy regular expression
   ↓
5. Applying additional rules for analysis
   ↓
6. Using regularity rules about attributes correlations
   ↓
7. Choosing template slots to be filled in
   ↓
8. Filling in default values in the obligatory template fields which remained empty
Additional Rules

Sometimes a diagnosis is given instead of organ description, e.g. "onychomycosis, tinea pedis". This information is captured at the fourth and fifth steps of the algorithm.

The word onychomycosis itself is not an attribute of any feature but it signals a fungal infection of the nail, so we can fill in the nail slot of our limbs template.

Example

"Крайници без отоци, запазени периферни пулсации, онихомикоза." (Lower limbs without oedema, preserved peripheral pulse, onychomycosis)
IE Steps

1. Selecting anatomic organ for processing
   ⇓
2. Finding relevant characteristics, features and related organs
   ⇓
3. Determining the scope of the expression for the chosen anatomic organ
   ⇓
4. Finding the most greedy regular expression
   ⇓
5. Applying additional rules for analysis
   ⇓
6. Using regularities rules about attributes correlations
   ⇓
7. Choosing template slots to be filled in
   ⇓
8. Filling in default values in the obligatory template fields which remained empty
We have observed features correlations. In case values for some of these features are missing we use the correlation rules to fill in the missing values.

**Example**

"Кожа - мугава с нормален тургор и еластичност." (Skin - dark with normal turgor and elasticity)
IE Steps

1. Selecting anatomic organ for processing

2. Finding relevant characteristics, features and related organs

3. Determining the scope of the expression for the chosen anatomic organ

4. Finding the most greedy regular expression

5. Applying additional rules for analysis

6. Using regularities rules about attributes correlations

7. Choosing template slots to be filled in

8. Filling in default values in the obligatory template fields which remained empty
Template slots

There are four manners to present the organs and their features:
1. General

![Diagram showing template slots for Limbs with parts and pulsations]
Template slots

There are four manners to present the organs and their features:
1. General

2. Explicit
Template slots

2. Explicit
Template slots

2. Explicit & 3. Partial
Template slots

2. Explicit & 3. Partial

4. By Diagnosis
IE Steps

1. Selecting anatomic organ for processing
2. Finding relevant characteristics, features and related organs
3. Determining the scope of the expression for the chosen anatomic organ
4. Finding the most greedy regular expression
5. Applying additional rules for analysis
6. Using regularities rules about attributes correlations
7. Choosing template slots to be filled in
8. Filling in default values in the obligatory template fields which remained empty
Template slots

- **Upper Limbs**
  - Arms: default
  - Peripheral Artery Pulsation: normally present

- **Lower Limbs**
  - Left leg
    - Lesg parts: slightly reduced
    - Peripheral Artery Pulsations: slightly reduced
  - Right leg
    - Lesg parts: default
    - Peripheral Artery Pulsations: normally present

- **Ankle**
  - Light oedema
## Evaluation I

<table>
<thead>
<tr>
<th>Feature</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>F-measure (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>88.89</td>
<td>90.00</td>
<td>89.44</td>
</tr>
<tr>
<td>#2</td>
<td>80.00</td>
<td>50.00</td>
<td>61.54</td>
</tr>
<tr>
<td>#3</td>
<td>98.28</td>
<td>96.67</td>
<td>97.47</td>
</tr>
<tr>
<td>#4</td>
<td>96.00</td>
<td>83.33</td>
<td>89.22</td>
</tr>
<tr>
<td>#5</td>
<td>95.65</td>
<td>73.82</td>
<td>81.33</td>
</tr>
<tr>
<td>#6</td>
<td>95.65</td>
<td>88.00</td>
<td>91.67</td>
</tr>
<tr>
<td>#7</td>
<td>94.94</td>
<td>90.36</td>
<td>92.59</td>
</tr>
<tr>
<td>#8</td>
<td>93.41</td>
<td>85.00</td>
<td>89.01</td>
</tr>
</tbody>
</table>

**Table:** IE precision, recall and f-measure evaluation (#1 - age; #2 - sex; #3 - diagnose; #4 - diabetes duration; #5 - skin; #6 - neck; #7 - thyroid gland; #8 - limbs).
## Evaluation II

<table>
<thead>
<tr>
<th>Template filled in at:</th>
<th>ankle</th>
<th>leg</th>
<th>peripheral artery</th>
<th>feet</th>
<th>skin</th>
<th>nails</th>
<th>others</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 4 by regex</td>
<td>150</td>
<td>148</td>
<td>199</td>
<td>6</td>
<td>9</td>
<td>14</td>
<td>18</td>
</tr>
<tr>
<td>Step 5 by extra rules</td>
<td>40</td>
<td>40</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Step 6 by correlations</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Step 8 by defaults</td>
<td>9</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>199</td>
<td>199</td>
<td>200</td>
<td>7</td>
<td>10</td>
<td>14</td>
<td>20</td>
</tr>
</tbody>
</table>

**Table**: IE performance steps 4-8 for limbs descriptions: number of PRs from where status statements are extracted at each step
Evaluation III

**Figure:** Extracted patient status for seven limbs attributes
In this article we described:

- A first try for automatic processing of medical documents in Bulgarian
- Evaluation of the performance for the extracted features from the anamnesis and patient status zones
- The role of explicitly-declared domain knowledge; it supports the information extraction algorithms by providing constraints and inference mechanisms
- Promising results which support the claim that the Information Extraction approach is helpful for the obtaining of specific medical statements which are described in the PR texts.

We plan to develop algorithms for discovering more complex relations and other dependencies.
Thanks for your attention!

Any questions?