Automatyczna analiza wypisów szpitalnych pacjentów chorych na cukrzycę

Małgorzata Marciniak, Agnieszka Mykowiecka

Plan prezentacji

- omówienie przetwarzanych tekstów,
- opis systemu SProUT,
- opis zadania ekstrakcji informacji z wypisów szpitalnych,
- wstępne wyniki,
- porównanie z zadaniem ekstrakcji informacji z tekstów mammograficznych,
- baza danych,
- jak wykorzystać system.

Osoby zaangażowane w projekt – spis alfabetyczny

- Roman Kuczerowski,
- Małgorzata Marciniak,
- Agnieszka Mykowiecka,
- Jakub Waszczuk.

Cel pracy

Wychwycenie w tekstach wypisów szpitalnych pacjentów diabetologicznych istotnych informacji służących do ustalenia przebiegu choroby i sposobu jej leczenia oraz zapisanie ich w bazie danych, aby ułatwić badania medyczne dotyczące cukrzycy.

Wypis szpitalny

- 1,5 – 2,5 strony tekstu w Wordzie ze stosunkowo niewielką liczbą błędów,
- teksty zawierają dane osobowe — wymagają więc wstępnego przetworzenia,
- część informacji zawarta jest w tabelach,
- interesujące nas informacje stanowią niewielki procent tekstu,
- przykładowy tekst.

Wyszukiwane informacje

- identyfikacja pobytu w szpitalu (numer wypisu, termin, przyczyna),
- dane dotyczace pacjenta (waga, wiek, płeć),
- dane dotyczące cukrzycy:
  - typ, charakterystyka, od kiedy:
    Cukrzyca typu 1, o chwiejnym przebiegu
    ...z wieloletnią cukrzycą typu 2
  - parametry cukrzycy: np. kwasica ketonowa (*aceton +*), stany niedocukrzenia,
- powikłania:
  Z późnych powikłań cukrzycy stwierdzono retinopatię prostą, nefropatię oraz polineuropatię obwodową.
  ...powikłana retinopatią prostą, neuropatią autonomiczną i obwodową

Wyszukiwane informacje — cd

- choroby autoimmunologiczne,
- ustalone leczenie (nazwy leków cukrzycowych, dawki insuliny):

  R 12 j Actrapid HM + 14 j. Insulatard HM
  P 9 j. Actrapid HM
  W 10 j. Insulatard HM
  Metformax 2x850 mg. na 30 min. przed posiłkiem
- ustalenie diety:

  Dieta cukrzycowa 1800 - 2000 kcal. 3 posiłki/dobę

  Dieta cukrzycowa 2000 kcal., 5-6 posiłków/dobę.

Wyszukiwane informacje — cd

- edukacja:
  Omówiono z chorą zasady diety, adaptacji dawek insuliny w zależności od różnych sytuacji (dodatkowa choroba, wysiłek fizyczny, niedocukrzenie, dodatkowe porcje jedzenia).
- przestrzeganie diety i samokontrola:
  Chory prowadzi samokontrole poziomu glukozy we krwi.,
  ...od pół roku nie prowadził samokontroli
- czy dokonano zmian schematu leczenia, dawek insuliny, diety:

  Utrzymano dotychczasowy system leczenia insuliną modyfikując jedynie dawki

  Zmieniono dotychczasowy system leczenia insuliną z dwóch wstrzyknięć na trzy

  Nieznacznie zmodyfikowano dietę.

SProUT

- **S**hallow Text **Pro**cessing with **U**nification and **T**yped Feature Structure, opracowany w DFKI Saarbrücken,
- dostosowany do współpracy z 11 językami w tym z polskim, wykorzystuje analizator morfologiczny Morfeusz,
- połączenie technik automatów skończonych z formalizmem unifikacyjnym,
- ścisła kontrola typów.

SProUT cd

Umożliwia:

- odwoływanie się do innych reguł gramatycznych: operator `@seek`,
- koreferencje zmiennych: `#z1`,
- tworzenie słownika dziedzinowego (*gazetteer*),
- definiowanie własnych funkcji dołączanych bezpośrednio do SProUT'a.

Przykładowa reguła

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

- Numer księgi głównej 11125/2006
- nr księgi głównej 12354
- Nr. księgi głównej 13578

Słownik zwany gazeterem

Monotard | GTYPE: gaz_oral | G_CONCEPT: monotard_t
Monotardu | GTYPE: gaz_oral | G_CONCEPT: monotard_t
Novorapidu | GTYPE: gaz_insulin | G_CONCEPT:novorapid_t
NovoRapidu | GTYPE: gaz_insulin | G_CONCEPT:novorapid_t
Novo Rapid | GTYPE: gaz_insulin | G_CONCEPT:novorapid_t
Novo Rapidu | GTYPE: gaz_insulin | G_CONCEPT:novorapid_t
neuropatie | GTYPE: gaz_comp | G_CONCEPT: neuropathy_t
Neuropatie | GTYPE: gaz_comp | G_CONCEPT: neuropathy_t
Neuropatia | GTYPE: gaz_comp | G_CONCEPT: neuropathy_t
neuropatią | GTYPE: gaz_comp | G_CONCEPT: neuropathy_t
neuropatię | GTYPE: gaz_comp | G_CONCEPT: neuropathy_t
retinopatii | GTYPE: gaz_comp | G_CONCEPT:retinopathy_t
retinopatią | GTYPE: gaz_comp | G_CONCEPT:retinopathy_t
retinopatię |GTYPE: gaz_comp | G_CONCEPT:retinopathy_t
retinopatia |GTYPE: gaz_comp | G_CONCEPT:retinopathy_t
niedocukrzenie | GTYPE: gaz_diab | G_CONCEPT:hipogl_t

Wykorzystanie słownika

```
insulina:>
    @seek(liczba) & [LICZ #jedn1]
    (token & [TYPE hyphen]
     @seek(liczba) & [LICZ #jedn2])?
    token & [SURFACE "j"]
    token & [SURFACE "."]?
    gazetteer & [GTYPE gaz_insulin, G_CONCEPT #rodzaj]
->insulin_treat_str & [I_TYPE #rodzaj,
                       DOSE_MIN #jedn1, DOSE_MAX #jedn2].
```

Leki Insulina:
R 18 j. Actrapid HM
P 7 - 10 j. Actrapid HM
W 22 j Mixtard 30 HM
Metformina 2 x 1 tabl. a 0,5 g.
Mononit 2 x 1 tabl. A 20 ,g. o 8.00 i 16.00

Reprezentacja informacji

Fragment hierarchii typów (SProUT):

insulin_treat_str := ∗avm∗ & [I_TYPE insulin_t,
DOSE_MIN string,
DOSE_MAX string].

Wynik rozpoznania napisu *9 j. Actrapid*:

$$\begin{bmatrix} \textit{insulin\_treat\_str} \\ \text{I\_TYPE} \quad \textit{actrapid\_t} \\ \text{DOSE\_MIN} \quad \text{“9”} \\ \text{DOSE\_MAX} \quad \textit{string} \end{bmatrix}$$

Fragment pliku tekstowego do zapisu w bazie danych:

DOSE_MAX:string || DOSE_MIN:34 || I_TYPE:actrapid_t

Problemy (te co zawsze)

- rozpoznawanie negacji:
  - *Po kilku dniach ustąpiła acetonuria.*
  - *Nie stwierdzono późnych powikłań cukrzycy pod postacią mikroangiopatii.*
- konieczność pisania reguł rozpoznajacych koordynację:
  - *retinopatię prostą oka lewego oraz proliferacyjna oka prawego z makulopatią w obu oczach*
  - *neuropatią autonomiczną oraz obwodową*
- niektóre wyrażenia mają różną interpretację w zależności od kontekstu, np. *mikroalbuminuria* może być zarówno powikłaniem: *wystąpiła mikroalbuminuria* jak i określeniem badania laboratoryjnego *Mikroalbuminuria: 25 mg/dobę*

Swoboda wypowiedzi

Informacje dotyczące edukacji, samokontroli oraz zmian w leczeniu mogą być wyrażone na bardzo wiele sposobów — konieczność rozpoznawania do kilkudziesięciu schematów:

- *Kontynuowano leczenie cukrzycy dotychczasowym systemem wielokrotnych wstrzyknięć modyfikując dawki oraz dietę.*
- *Utrzymano dotychczasowy system wielokrotnych wstrzyknięć insuliny korygując dawki.*
- *Zmieniono system leczenia – podano pacjentce insulinę w dwóch wstrzyknięciach na dobę.*
- *Utrzymano dotychczasowy system wielokrotnych wstrzyknięć insuliny zmieniając dawki poszczególnych insulin, stosując dawki interwencyjne oraz modyfikując system dietetyczny.*
- *Kontynuowano dotychczasowy schemat leczenia.*
- *Kontynuowano wcześniejsze leczenie hipotensyjne. !!!!*

Wstępne wyniki

Dla raportów:

	liczba	precyzja	pełność
typ cukrzycy	51	100	98,04
cukrzyca niewyrównana	22	100	86,36
komplikacje	44	93,62	100

Dla fraz:

	liczba	precyzja	pełność
typ cukrzycy	53	100	98,11
cukrzyca niewyrównana	30	100	76,67
komplikacje	91	96,59	93,41

Porównanie zadań ekstrakcji

Mammografia	**Diabetologia**
teksty bardzo krótkie (kilka linijek) pisane przeważnie równoważnikami zdań	teksty długie (około 2 stron w Wordzie), zawierają wiele tabel
teksty niestaranne – oryginalne dane zawierały sporo błędów pisowni, które trzeba było poprawić	teksty stanowią dokumentację szpitalną i są najczęściej poprawne
niemal cały tekst zawiera istotne informacje	wyszukiwana informacja stanowi niewielki procent tekstu
niemal wszystkie wyszukiwane informacje to proste frazy, przy czym ta sama informacja może być wyrażona na kilka sposobów	większość informacji wyrażona jest przez bardzo proste schematy, jednak część wymaga rozpoznania kilkudziesięciu możliwych wariantów

Porównanie zadań ekstrakcji – cd.

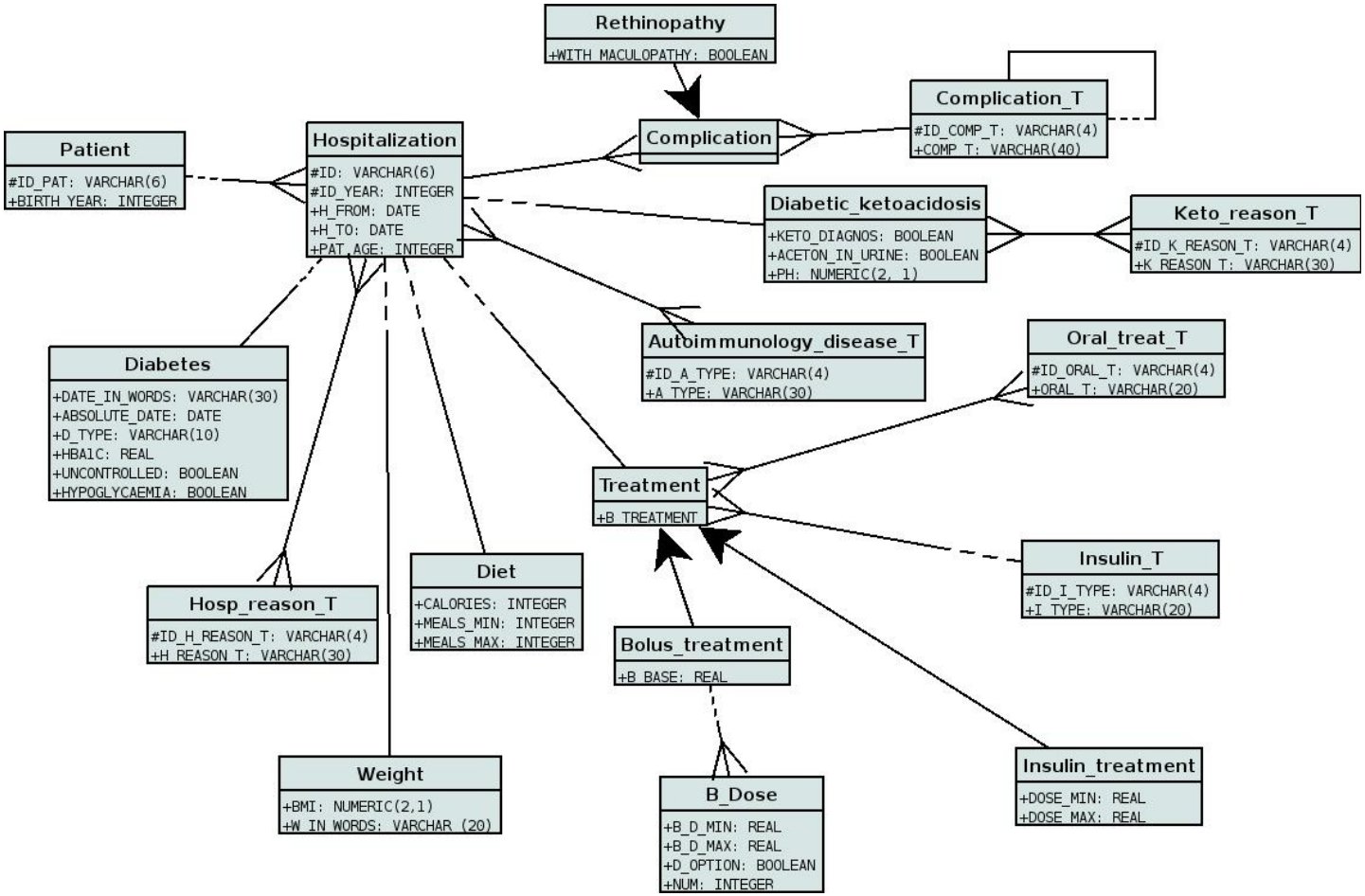
Mammografia	**Diabetologia**
informacje się nie powtarzają, wszystkie wyszukane należy uwzględnić przy dalszym przetwarzaniu	istotne informacje się powtarzają, a część wyszukanych informacji należy pominąć przy dalszym przetwarzaniu
wykorzystywane są informacje morfologiczne, m.in. do wyznaczania granicy fraz	niemal całkiem zrezygnowano z wykorzystywania informacji morfologicznych
dalsze przetwarzanie wyszukanych informacji jest skomplikowane i wymaga metod heurystycznych	wyszukane informacje wprowadzane są do bazy danych przy użyciu nieskomplikowanych algorytmów

Baza danych

Wyniki uzyskane z systemu SProUT są przetwarzane na zawartość relacyjnej bazy danych.

- dane wejściowe:
  - hierarchia licząca ponad 200 typów,
  - wyniki gramatyk SProUTa,
  - dodatkowe zasady interpretacyjne, np. uwzględniaj leki występujące po atrybucie 'epikryza',
- dane wyjściowe:
  - około 25 tabel bazy relacyjnej (PostgreSQL 1.6.2)

Schemat bazy danych



Podsumowanie etapów przetwarzania danych

1. dane oryginalne – zbiór plików MSWord
   ↓ usuwanie danych identyfikacyjnych – skrypt Perl
2. dane z kodami identyfikacyjnymi pacjentów – zbiór plików Word
   ↓ skrypt Perl
3. jeden plik tekstowy w kodowaniu UTF-8
   ↓ wyszukanie interesujących informacji (SProUT)
4. plik XML
   ↓ usunięcie niepotrzebnych informacji i etykiet – skrypt Perl
5. plik tekstowy z wybranymi atrybutami
   ↓ program C++
6. baza danych
   ↓ zapytania SQL

   informacje o pacjentach

Przykładowe informacje o zawartości bazy danych

- 278 hospitalizacji,
- typy cukrzycy: I - 88, II - 186, inne - 4,
- typy komplikacji:

coronary_disease	90
autonomic_neuropathy	14
nephropathy_t	41
cerebrovascular_disease	3
preproliferative_retino	7
diabetic_foot	8
retinopathy_t	30
microalbuminury_t	14
peripheral_vascular_disease	17
nonproliferative_retino	85
peripheral_polyneuropathy	44
macroangiopathy_t	34
proliferative_retino	15

Planowane wykorzystanie systemu

- Sprawdzenie czy w wypisach znajdują się informacje, które powinny być obowiązkowo wpisane, np. od kiedy pacjent choruje na cukrzycę.

- ustalenie zależności występowania pewnych czynników:
  - ustalenie jaki procent chorych na cukrzycy typu 2 z poziomem $\mathrm{HbA}_{1C} > 7.5$ jest leczonych insuliną,
  - jaki procent chorych z poziomem kreatyniny powyżej 1,3 cierpi na powikłania o charakterze mikroangiopatii (nefropatia, retinopatia)
  - zależność pomiędzy poziomem kreatyniny a występowaniem nadciśnienia tętniczego.

Kierunki dalszych prac

- zdefiniowanie reguł rozpoznających nowe frazy reprezentujące konkretne typy informacji na podstawie już oznaczonych danych, np. dla rozpoznań i komplikacji