Анонимизация на клинична информация за пациенти

ДИПЛОМНА РАБОТА НА МАРИН НОЖЧЕВ
БИО- И МЕДИЦИНСКА ИНФОРМАТИКА

ДИПЛОМЕН РЪКОВОДИТЕЛ: **ГЛ. АС. Д-Р КАЛИН ГЕОРГИЕВ**

Накратко

Основната цел на дипломната работа е създаване на Софтуер за **премахване на лични данни** от свободни клинични текстове с цел **улесняване на споделянето на информация** между **медицински заведения** и **научни институции**

Какво е анонимизация на медицински данни?

КЛИНИЧЕН ТЕКСТ

ADMISSION DATE :02/28/1999
DOB :9/10/67

Mrs. Given is a 31 year old female with a recent history of pneumonia as well as polysubstance abuse , depression , multiple suicide attempts , who actually came to the emergency room after visiting her 51-year-old boyfriend.At about an hour and a half after the patient ....This was witnessed by her 12 year old daughter , who called the EMT 's .

КЛИНИЧЕН ТЕКСТ БЕЗ ЛИЧНИ ДАННИ

ADMISSION DATE :[[ДАТА]]
DOB : [[ДАТА]]

[[ИМЕ]] is a [[ВЪЗРАСТ]] female with a recent history of pneumonia as well as polysubstance abuse , depression , multiple suicide attempts , who actually came to the emergency room after visiting her [[ВЪЗРАСТ]] boyfriend.At about an hour and a half after the patient ....This was witnessed by her [[ВЪЗРАСТ]] daughter , who called the EMT 's .

Приложение на анонимизираната информация

- Данни от лечението на големи групи пациенти са достъпни за анализ
- Чрез data mining се откриват доказателства за
  - Ефективност на лекарства
  - Методи на лечение
  - Управленски практики на медицински заведения

Аспекти на анонимизацията

АНОНИМИЗАЦИЯ НА СТРУКТУРИРАНИ ДАННИ

```
<?xml version="1.0" encoding="UTF-8"?>
<PATIENT>
 <PHI TYPE="HOSPITAL">DH</PHI>
 <PHI TYPE="PATIENT">ED FRANCYIE KOTEVERGE</PHI>
 <UNIT_NUMBER>870-79-47</<UNIT_NUMBER>
</PATIENT>
```

… НА НЕСТРУКТУРИРАНИ ДАННИ

During the winter and early parts of 1993 , the patient noticed increasing fatigue , symptoms of breathlessness and recurrent substernal chest pressure which finally led to admission to Xas Tupalmsmodral Hospital on 11-11-93 with documentation of a subendocardial myocardial infarction and a blood pressure of **210/108** .

Критерии за качество

- Неформален: Каква част от личните данни са заличени?
- Тривиална формализация на критерия е неефективна
- Използват се мерки от анализа на текст

$$precision: P = \frac{TP}{TP+FP}$$

$$recall: R = \frac{TP}{TP+FN}$$

$$F - score: F_\beta = \frac{(1+\beta^2)P \cdot R}{\beta^2 P + R}$$

Обзор на подходите за анонимизация

МАШИННО САМООБУЧЕНИЕ

Подход: Текстът се моделира като процес, при който това, дали текущата текстова единица е лични данни, е условна вероятност, която зависи от предишните състояния.

Реализации: Скрити Марковски модели, Марковски модели с максимална ентропия, Conditional Random Fields

АНОНИМИЗАЦИЯ, БАЗИРАНА НА ПРАВИЛА

Подход: Дефинират се правила, които свързват даден контекст с наличието на лични данни

Реализации: Регулярни изрази, Граматики

Обзор на подходите за анонимизация: предимства и недостатъци

МАШИННО САМООБУЧЕНИЕ

+ изискват по-малко ръчно адаптиране към конкретен сборник от текстове

- изискват голямо количество примерни текстове

- не позволяват фина ръчна настройка на резултата

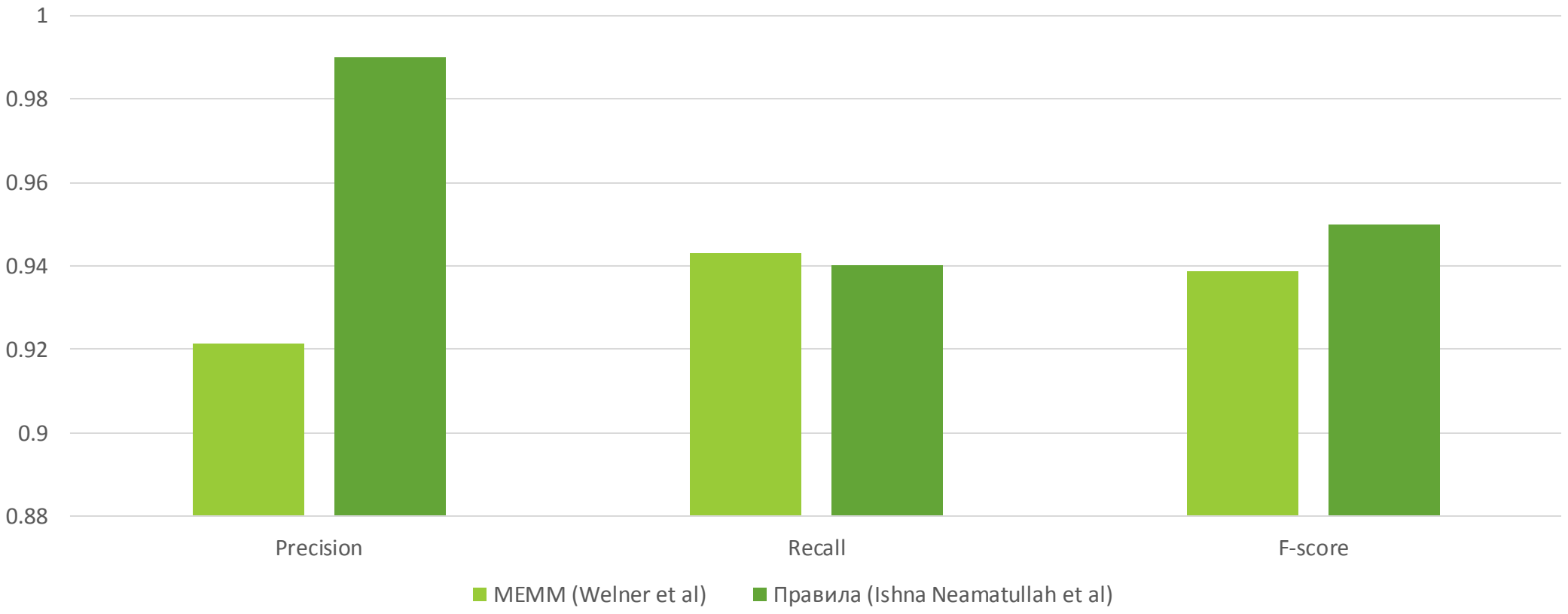
АНОНИМИЗАЦИЯ, БАЗИРАНА НА ПРАВИЛА

- изискват ръчно адаптиране към конкретен сборник от текстове

+ изискват малко на брой примерни текстове

+ експерти в областта лесно могат да подобряват алгоритъма

Обзор на подходите за анонимизация: публикувани резултати



Архитектура на софтуера за анонимизация

- Алгоритъм за анонимизация: базиран на компоненти за разпознаване на именувани обекти и правила
- Платформа за анализ на текст: GATE 7.1 (University of Sheffield)
- Компоненти за анализ на текст: базирани на ANNIE (University of Sheffield)
- Платформа на изпълнимия код: Oracle Java 7
- Вход: XML или текстови файлове
- Изход: Анотирани файлове в XML формат или анонимизирани текстови файлове

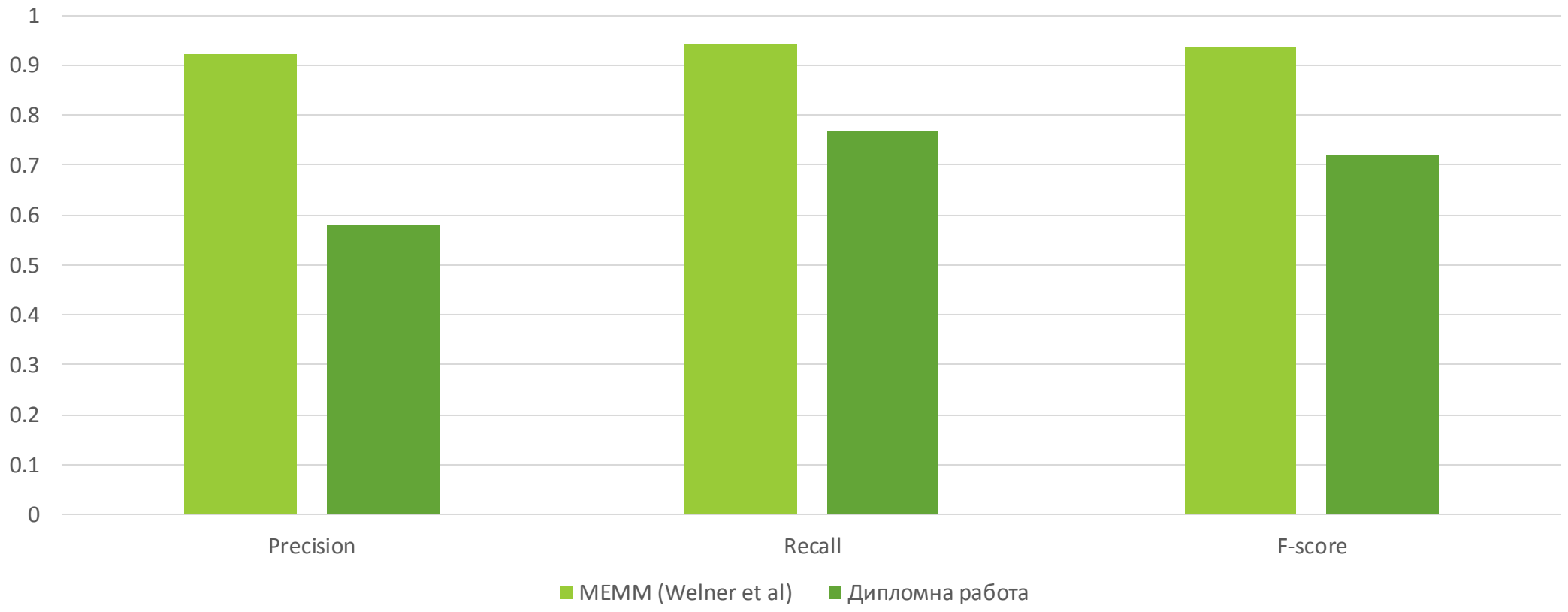
Архитектура: възможности за интеграция

- Като Java библиотека или OSGi компонент
- Чрез JNI в Windows и Linux приложения, включително скриптови езици като Python
- Като GATE Processing Resource за приложения базирани на платформата GATE.

Възможности за разширение

- Поддръжка на български език чрез готови компоненти за платформата GATE
  - Идентификатор на части на речта, базиран на BulTreeBank (д-р К. Симов) и LingPipe
  - Разпознаване на именувани обекти, базирано на CRF (д-р Г. Георгиев, д-р П. Наков, К. Ганчев)
- Пакетиране на софтуера като Уеб услуга (Web Service)
  - Позволява интегрирането във всички архитектури, вкл. ERP приложения

Сравнение на софтуера с публикувано state-of-the-art решение



Въпроси?

БЛАГОДАРЯ ЗА ВНИМАНИЕТО!

Резервна информация

Компоненти в детайли