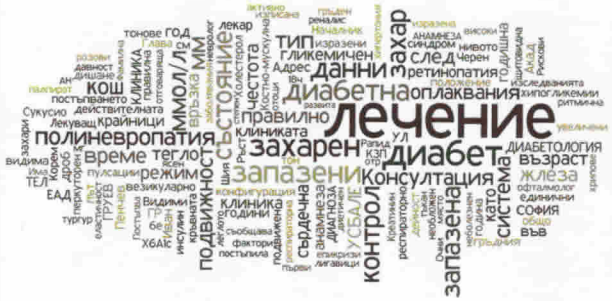
Ефективно търсене на концептуални шаблони с приложения в медицинската информатика

ЕВТИМА



Фонд
"Научни изследвания"
Конкурс "Идеи-2008"
Договор ДО-02-292
2009 – 2011

http://www.lml.bas.bg/evtima

Цели на Проекта

Участници

Резултати

План за работа

Извличане на факти от записи на пациенти

↓

Построяване на формално вътрешно представяне

↓

Търсене на закономерности

Ефективно търсене на концептуални шаблони с приложения в медицинската информатика

ЕВТИМА

Фонд „Научни изследвания“

Конкурс „Идеи-2008“

Договор ДО-02-292

2009 – 2011

http://www.lml.bas.bg/evtima

Актуалност на задачата

В целия свят се наблюдава нарастващ интерес към автоматичната обработка на текстове в областта на биомедицината. Има различни причини за това:

• *изключително бързото нарастване на обема информация в текстов формат* (например през 2005 г. в архива на Medline е постъпвала ежедневно библиографска информация средно за 1775 нови научни публикации),

• *повишени изисквания към качеството на медицинско обслужване на пациента,*

• *стремеж за подобряване на здравния мениджмънт чрез автоматично извличане на факти от медицински документи* и др.

В тази сравнително нова област компютърната лингвистика среща редица специфични проблеми:

• *Огромно терминологично богатство*, което затруднява създаването на специализирани речници, независимо от ограниченията на стила и жанра,

• Необходимост *текстовите единици да се интерпретират като понятия и отношения* от предварително зададен концептуален (смислов) модел на областта – а този модел също трябва да се създаде ръчно,

• Липса на *стандарти за номенклатурите и класификациите,* както и на *добри практики за тестване на резултатите, за пренос на разработки от друг естествен език* и т.н.

Цели на Проекта

В тригодишен период ЕВТИМА изследва проблемите на ефективно търсене на концептуални шаблони чрез методи за бързо намиране на повтарящи се отношения между понятията в голям масив от факти. Проектът използва оригинална идея за двустепенна обработка на концептуална информация. На предварителната фаза ще се пресметнат всички възможни концептуални шаблони в предметната област и ще се компресират в минимален краен автомат с маркери на заключителните състояния. Обработката по време на изпълнение ще се сведе до търсене на пътища в автомата, което се извършва изключително бързо. Експерименталните дейности в проекта са насочени към приложение на описания метод над факти, извлечени автоматично от текста на болнични записи на български език.

Целта на първия етап на ЕВТИМА е да се положат основите на експеримента, като се създаде среда за автоматична текстообработка на медицински записи на български език. В процеса на работа се открояват и други конкретни задачи, които осигуряват постигането на главните цели на проекта и продължение на дейностите след 2011 година:

- Създаване на езикови ресурси и разработка на специфични софтуерни модули за правописна корекция на медицински текст;
- Проектиране на инфраструктура за поддръжка на българска медицинска терминология, с възможност за връзки към многоезикови ресурси;
- Натрупване на анотирани учебни корпуси и структурирани тестови данни за създаването на технологии за автоматична обработка на български медицински текст.

Автоматично извличане на факти от биомедицински текстове

Софтуерните системи за автоматично откриване на факти обикновено работят чрез последователни стъпки за текстообработка:

• Разделяне на входния текст на думи (низове от букви), числени низове, пунктуационни маркери и т.н.;

• Намиране на думи, които съответстват на термини и на уникални наименования (например имена на хора и лекарства), на дати, важни съкращения и др. под. Системите разпознават намерените низове благодарение на предварително създадени речници (лексикони) на термини, думи, имена и т. н. На този етап следва да се разпознаят и евентуално коригират правописните грешки във входния текст;

• Разпознаване на референтните цитирания на един обект – например, термин може да е заместен с местоимение или синоним;

• Разпознаване на отношенията между идентифицираните обекти. Често тези отношения не са изразени явно, а се подразбират, тъй като документът е написан от човек-експерт и е адресиран към друг човек-експерт, който познава областта и смисловите връзки между понятията. Софтуерната система "разбира" текста, като автоматично свързва обектите с явно наименовани отношения. Явното наименоване е възможно благодарение на предварително зададени концептуални модели на понятията в областта и връзките между тях;

• Запълване на вътрешна структура, обикновено таблица, с намерените понятия и отношения в текста.

Точността на извличане се измерва с поредица от тестове над непознати за системата документи. Най-често системите са настроени да намират важни понятия, събития и връзки, без да обработват целия входен текст.

Създаването на електронни ресурси в медицината – лексикони и концептуални модели – е едно изследователско и приложно предизвикателство, поради голямото количество понятия и връзки между тях.

Участници

В проекта ЕВТИМА участват две организации:

Секция за лингвистично моделиране (СЛМ) на Института по паралелна обработка на информацията на Българската академия на науките (ИПОИ-БАН) – водещо българско звено в научните и приложни изследвания по компютърна лингвистика и езикови технологии, изкуствен интелект и обработката на знания (http://www.lml.bas.bg). В многобройните си международни проекти СЛМ е натрупала значителен опит при създаване на разнообразни прототипи за обработка на естествен език, както и лингвистични ресурси на български език.

Отделение по медицинска информатика в Университетската специализирана болница за активно лечение по ендокринология "Акад. Ив. Пенчев" (УСБАЛЕ) към Медицинския университет – София. УСБАЛЕ е пионер на българската медицинска информатика и един от основателите на българската медицинска мрежа MedicalNet-BG. Болничната информационна система понастоящем обхваща над 150 000 записа на пациенти, които отговарят на изискванията на MBDS (Minimal Basic Data Set).

Към основния колектив са привлечени експерти по анотация на медицински текстове, представяне на знанията, разработка на интелигентни системи. Две трети от сътрудниците на ЕВТИМА са докторанти, пост-докторанти или учени под 40 години.

План за работа

През първия етап на проекта основните дейности са:

- Проучване, проектиране и детайлно планиране на разработката с оглед необходимите потребителски услуги;
- Провеждане на теоретични изследвания, свързани с тематиката на проекта;
- Създаване на начални версии на софтуерните прототипи и оценка на работата им в експериментални условия;
- Организиране на научни мероприятия и участие в конференции и семинари с цел разпространение на получените резултати.

Дейностите през втория етап включват задълбочаване на изследванията и разширения на експерименталната разработка.

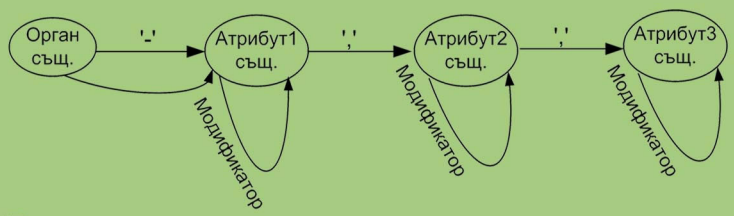
Автоматична обработка на болнични записи на български език

Макар и структурирани в отделни зони, текстовете на записите са необичайно трудни за автоматична текстобработка. Преди всичко те съдържат много числови стойности и думи, които не се срещат в речник на общата българска лексика. Долната таблица характеризира низовете от символи, срещнати в учебния корпус от болнични записи. Значителният брой непознати думи изисква съставяне на отделен речник на медицинската терминология.

	Общо низове	Неезикови единици		Непознати думи		Разпознати словоформи
		Брой	%	*Брой*	%	
Учебен корпус	65 600	6 680	10,2%	13 370	20,3%	45 550

В текста се съдържат термини на български и латински език, като латинските често са транскрибирани на кирилица. Някои термини се записват с различни съкращения. Пълните, завършени изречения се срещат рядко, което не позволява да се приложат традиционните процедури за автоматичен синтактичен анализ. Текстът е съставен предимно от отделни декларативни фрази, разделени с пунктуационни символи. Обсъждането на важна тема почти винаги започва в ново изречение, но в рамките на отделните зони (анамнеза, статус, обсъждане) не се практикува минаване на нов ред. От гледна точка на компютърната лингвистика, изброените специфични особености правят текста на болничните записи много труден обект за автоматична обработка.

Извличането на факти за пациенти в ЕВТИМА се осъществява чрез техники за частичен анализ на текста. Създадени са линейни правила за намиране и разпознаване на термини и ключови фразови конструкции, които ги съдържат. На практика се приема, че споменаването на дадена езикова конструкция в текста е положително твърдение за наличие на изказания факт. Чрез правила се обработват и редица форми на отрицание, които показват отсъствие или редуциране на даден факт или признак.



Графично-зададено морфосинтактично правило за откриване на типични фрази в болнични записи

Пример:
Крайници – без отоци, запазени периферни пулсации, онихомикоза
орган атрибут 1 атрибут 2 атрибут 3

Резултати

Създадени са софтуерни процедури за обработка на текста на болнични записи. Съставят се необходимите компютърни речници и концептуални модели, описващи предметната област. Терминологичните колекции на български език следват наличните стандартизирани болнични номенклатури МКБ-9 и МКБ-10. Концептуалните модели са много важни при анализа на текста, тъй като декларираните в тях отношения между понятията позволяват да се разграничат различните теми в записа на пациента. По този начин системата разбира как да фрагментира текста в отделни структурни единици. При създаване на концептуални модели ЕВТИМА ползва публични източници за английски език в Интернет, включително най-разпространения в компютърната лингвистика ресурс UMLS (Unified Medical Language System).

Разработените софтуерни модули са свързани в прототипна система, която извлича автоматично сведения за диагнозата, продължителността на заболяването и статуса на пациентите. Намерените данни се внасят в таблица, от която лесно се произвежда вътрешно формално-логическо представяне на фактите. През юни 2010 г. учебният корпус на системата се състои от 197 анонимизирани записа, а тестовете се провеждат над 1000 анонимизирани записа. Коректността на извличането се оценява чрез измерване на две типични характеристики:

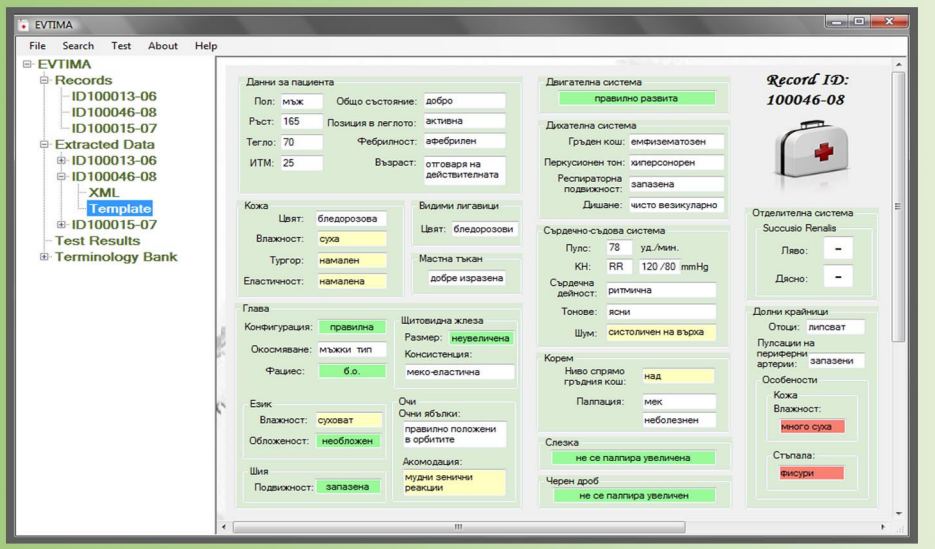
- ***точност*** *(precision)* – процент коректно разпознати стойности от всички *извлечени* стойности;
- ***покриваемост*** *(recall)* – процент коректно разпознати стойности от всички *налични* в тестовия корпус.

Анализът на текста се основава върху емпирично изведени правила за линейна съчетаемост на думите. По-сложните езикови конструкции изискват допълнителна обработка. Долната таблица показва постигнатите до момента успехи при автоматично разпознаване и извличане на факти за диагнозата, давността на заболяването и статуса на диабетици.

	Диагноза	Давност на диабета	Състояние на кожата	Състояние на шията	Състояние щит. жлеза	Състояние крайници
Точност	98,28%	96,00%	95,65%	95,65%	94,94%	93,41%
Покриваемост	96,67%	83,33%	73,82%	88%	90,36%	85,00%

Резултати

Основна цел на процедурите за автоматично структуриране на информацията е да се разпознаят и класифицират споменатите в текста характеристики на пациента. При описание на естествен език обаче се използват множество думи в различни лексикални варианти, както и свободни терминологични парафрази. Само за статуса на кожата в епикризите от учебния корпус са намерени 93 характеристики. Създадената прототипна система автоматично класифицира научените думи и фрази в състояния на пациента: *добро, леко увредено, увредено* и *силно увредено*. В прототипа е заложено и знание за стойностите по подразбиране, с които могат да се запълват пропуснатите в текста описания. Така системата успява да запълни задължителните полета от статуса на пациента и да "нормализира" разнообразните състояния, като ги сведе към предварително зададени усреднени стойности.



Структуриране на данни за статуса на пациента, автоматично извлечени от болнични записи

За контакти:

доц. дмн Галя Ангелова, БАН, http://www.lml.bas.bg/~galja

доц. д-р Димитър Чаръкчиев, УСБАЛЕ, http://www.medicalnet-bg.org/dimitar

Дата на издаване: юни 2010 г.