БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ
ИНСТИТУТ ПО ИНФОРМАЦИОННИ
И КОМУНИКАЦИОННИ ТЕХНОЛОГИИ

Ивелина Мирчева Николова

ПРИЛОЖЕНИЕ НА ОБРАБОТКАТА НА ЕСТЕСТВЕН ЕЗИК ЗА ИЗГРАЖДАНЕ НА СЕМАНТИЧНИ СИСТЕМИ

Д И С Е Р Т А Ц И Я

за присъждане на

образователната и научна степен „Доктор“

Научна специалност: 01.01.12 „Информатика“

Професионално направление: „Информатика и компютърни науки”

Научен ръководител: проф. д.м.н. Галя Ангелова

Научен консултант: доц. д-р Димитър Чаръкчиев

София, 2014 г.

Съдържание

Благодарности **5**

Увод **6**

Актуалност на темата 6

Цели и задачи на дисертационния труд 8

Структура на дисертационния труд 10

1 Обзор на основните резултати в областта **11**

1.1 Характеристики на семантичните системи 11

1.2 Разпознаване на парафрази 15

1.3 Извличане на понятия и релации между тях 18

1.4 Автоматично структуриране на описания от пациентски записи . 22

2 Приложение на OEE за създаване на концептуални модели на предметна област **24**

2.1 Въведение 24

2.2 Разпознаване на синоними и парафрази на понятията от наличната терминологична банка 29

2.2.1 Необходимост и приложение на разпознаването на парафрази 29

2.2.2 Дефиниция на задачата 30

2.2.3 Метод 30

2.2.4 Ресурси 33

2.2.5 Моделиране на фразова структура и правила за разпознаване на парафрази 36

2.2.6 Експерименти и резултати 40

2.2.7 Резюме 41
2.3 Разпознаване на релации между медицински понятия чрез обработка на дефиниции от UMLS 43
2.3.1 Въведение 43
2.3.2 Дефиниция на задачата 44
2.3.3 Ресурси 45
2.3.4 Метод 47
2.3.5 Дискусия 56
2.3.6 Резюме 58

3 Структуриране на текстови описания в биомедицината 59
3.1 Извличане на симптоми на диабет от пациентски записи 61
3.1.1 Въведение 61
3.1.2 Дефиниция на задачата 61
3.1.3 Ресурси 62
3.1.4 Методи 66
3.1.5 Оценяване 71
3.1.6 Резюме 75
3.2 Разпознаване на събития и тяхната темпоралност в медицински текстове 76
3.2.1 Въведение 76
3.2.2 Ресурси 78
3.2.3 Метод 79
3.2.4 Експерименти и резултати 86
3.2.5 Дискусия и анализ на грешките 89
3.2.6 Резюме 90

4 Интеграция в прототипи 91
4.1 Вграждане на МС за класификация на атрибути в среда за автоматично откриване на потенциални диабетици в хранилище с големи данни 91
4.1.1 Въведение 91
4.1.2 Дефиниция на задачата 92

4.1.3 Входни данни 93
4.1.4 Експерименти и резултати 93
4.1.5 Резюме 98

Заключение **100**
Основни научни и научно-приложни приноси 103
Списък на публикациите по дисертацията 106
Апробация на резултатите 110
Участие в национални и международни проекти 111
Декларация за оригиналност на резултатите 112

Използвани термини, съкращения и означения **112**

Списък на фигурите **119**

Списък на таблиците **121**

Библиография **122**

Благодарности

Бих искала да изкажа най-искрени благодарности на хората, които бяха до мен по този 4-годишен път. Тези, които споделиха с мен време в работа и трупане на ценен опит - Светла Бойчева, Димитър Чаръкчиев, Елена Паскалева, Кевин Коен и Ирина Темникова. С особено внимание ще откроя моя научен ръководител Галя Ангелова, която винаги е заставала до мен с пълната си подкрепа и окуражаващи думи. Благодаря на семейството и приятелите ми, че ми дават любовта си и правят живота ми цветен.

Бих искала също така да изкажа благодарност за финансовата подкрепа, без която това изследване не би могло да се случи:

Проект DO 02-292/Декември 2008 "Ефективно търсене на концептуални шаблони с приложения в медицинската информатика (ЕВТИМА)", финансиран от Националния фонд за научни изследвания през 2009-2012.

Проект BG051PO001-3.3.04/40: Изграждане на висококвалифицирани млади изследователи по съвременни информационни технологии за оптимизация, разпознаване на образи и подпомагане вземането на решения.

Проект No 216130 / 1.05.2010 - 31.07.2011: Сигурност на пациента чрез интелигентни процедури в лечението (PSIP+)

Проект BG051PO001/3.3-05 „НАУКА И БИЗНЕС" на МОМН - лична стипендия за едномесечно обучение в Денвър, Колорадо - University of Colorado School of Medicine. Работа върху извличане на събития от медицински епикризи на пациенти, диагностицирани с различни заболявания.

Проект No. 316087 "Съвременните пресмятания в полза на иновацията (AComIn)", финансиран по FP7 Capacity Programme на Европейската Комисия през 2012–2016.

Увод

Актуалност на темата

Масовото навлизане на информационните технологии в съвременния живот е съпътствано от неминуемо бързо увеличаване на обема на дигитална информация. Ежедневно се натрупват записи от работата на информационни системи, системи за диагностика и наблюдение, медиите ползват електронни канали за разпространение на новини, историческото наследство и изкуството се дигитализират, създава се електронна мултимедия. Налични са големи количества електронни записи с разнообразни източници, категория, съдържание и формат. На личните си компютри и мобилни устройства ние също създаваме записи, запазваме мултимедийни документи. Така естествено се появява нуждата от ефективни технологии за съхраняване, класифициране, обработване, синтезиране на електронна информация и за търсене в нея, както от хора, така и от компютри.

Семантичните системи (СС) са именно такава технология. Те съчетават предимствата на база данни и интелигентните подходи за обработване на информацията. За разлика от конвенционалните системи за търсене в общ контекст и в Интернет, които съхраняват и индексират информация от хетерогенни източници, СС обработват тази информация чрез интелигенти алгоритми за семантична анотация, представят я във вид на семантични бази данни, свързват обектите в текста с точно определени съществуващи обекти в онтологичното знание на системата за света, и също така предоставят достъп до данните чрез интелигенти търсещи машини (intelligent content-based search). Те индексират както интересните обекти, така и връзките между тях, което позволява освен търсене на понятия и контекстуализирано търсене. Такива системи са ценни за

анализ и разглеждане на големи масиви от данни, защото дават възможност за бързо извличане на информация за дадени обекти в полу-структуриран вид, който е удобен и за последваща машинна обработка. Обикновено се реализират в ограничена предметна област, което подобрява точността на вложените алгоритми.

В тази дисертация се разглеждат конкретно семантичните системи, в които се предполага извличане на информация от текст – семантични системи, обработващи текст (ССТ). В тях информацията постъпва като свободен или частично структуриран текст, който с помощта на техники за обработка на естествен език (ОЕЕ) се структурира спрямо даден концептуален модел и в последствие се индексира и записва в семантични бази данни.

Техниките за ОЕЕ са от ключово значение за ССТ. Те представляват междинен слой, в който входните данни на системата се структурират от свободен текст до понятия и връзки спрямо модела на предметната област. Целта на тази дисертация е да изследва приложението на някои техники за ОЕЕ в ССТ в областта на биомедицината за български и английски език. Също така да се провери наличността на ресурси за реализирането на интегрирани семантични системи, работещи над български език в тази област. Да се очертаят насоките за развитието им с цел създаването на висок клас семантични системи, които да организират съдържанието на текста, понятията в него и отношенията между тези понятия с висока точност, да работят прозрачно за потребителите и да им служат по интуитивен начин. Наличието на подобни системи би помогнало да се открият зависимости, извлечени от текстовите описания на медицинските документи чрез интелигентните алгоритми за анализ на текст, и те да се предоставят на специалистите, от които зависи да вземат решения с цел подобряване на здравните политики и мениджмънта в българската здравна система.

Цели и задачи на дисертационния труд

Настоящата работа бе вдъхновена от стартиралите в ИИКТ през 2010 г. проекти за автоматичен анализ на клиничен текст на български език с идеята, че най-после и у нас ще започне натрупване на научен капацитет, знания, опит, ресурси и софтуер за обработка на медицински записи. Принципните цели на дисертацията са както следва:

- Да се изследват начини за създаване и разширяване на формализирани модели в областта на медицината чрез автоматично извличане на концептуални единици от текстови дефиниции.
- Да се изследват възможностите за създаване на декларативни описания на предметната област на български език (беден на електронни медицински ресурси) чрез използване на по-богатите ресурси налични на английски език.
- Да се разработят подходи за структуриране на информация за състоянието на пациента чрез извличане на характеристики от свободен текст на български език и запълване на шаблони.

Пред автора бяха поставени следните конкретни цели и задачи:

1. Да се изследват техниките за автоматично разпознаване на парафрази в компютърната лингвистика. Да се изследват езиковите структури в пациентски записи на български език, да се подберат подходящи методи и да се разработи прототип за автоматично разпознаване на парафрази, като средство за атакуване на вариативността на естествения език при извличане на информация за състоянието на пациента.
2. Да се изследват големите медицински онтологии, интегрирани в UMLS, относно наличието на декларативно-представени релации между понятията. За целта да се разучи и използва софтуерът за извличане на справки от UMLS.
3. Да се изследват техниките за автоматично извличане на релации между понятия от текст, с цел обогатяване на декларативен концептуален модел на предметната област. Да се реализира прототип за извличане на релации, значими за медицински приложения.

4. При реализирането на прототипите за извличане на релации и автоматично разпознаване на парафрази, да се експериментира както с базирани на правила подходи, така и със статистически методи за обработка на естествен език. Да се разработи хибриден подход, при който статистическите техники подпомагат и извличане на правила за анализ от изходния текст на първичните документи.
5. Да се изследват техники за извличане на информация, които се използват при автоматичен анализ на биомедицински текстове. Да се разработят прототипни компоненти за структуриране на информация за състоянието на пациента спрямо няколко важни и типични характеристики. Да се извършат експерименти с подходи, базирани на правила и с методи за машинно самообучение.
6. Да се демонстрира качеството на разработените прототипи чрез интегрирането им в семантични системи, използващи автоматична обработка на естествен език. Тестването на приложимостта и успеваемостта да се извършва над реални данни, по възможност на български език.

Структура на дисертационния труд

Дисертацията е структурирана в увод, 4 глави и заключение, и заема 129 страници. В увода е направено кратко въведение в темата и са поставени целите и задачите на докторанта. Първата от 4-те основни глави представя обзор на основните резултати в областта. Втората глава описва изследванията на докторанта върху приложения за обработка на естествен език при създаване на концептуален модел на дадена предметна област и по-конкретно в биомедицината. Тази глава представя методи за извличане на парафрази и релации между понятия. В третата глава са представени подходи за извличане на структурирани описания от медицински текстове на български и на английски език. Четвърта глава демонстрира интеграция на описаните по-горе методи в прототипи. В заключението е направено резюме на представените подходи и резултати, както и обобщени изводи в резултат от анализите. Представен е списък с публикациите по дисертацията и събития и проекти, при които са апробирани резултатите. Описани са приносите на докторанта. Представена е декларация за оригиналност на резултатите. Следват списък на фигурите, списък на таблиците и библиографска справка на използваната литература. Цитирани са 67 източника.

Глава 1

Обзор на основните резултати в областта

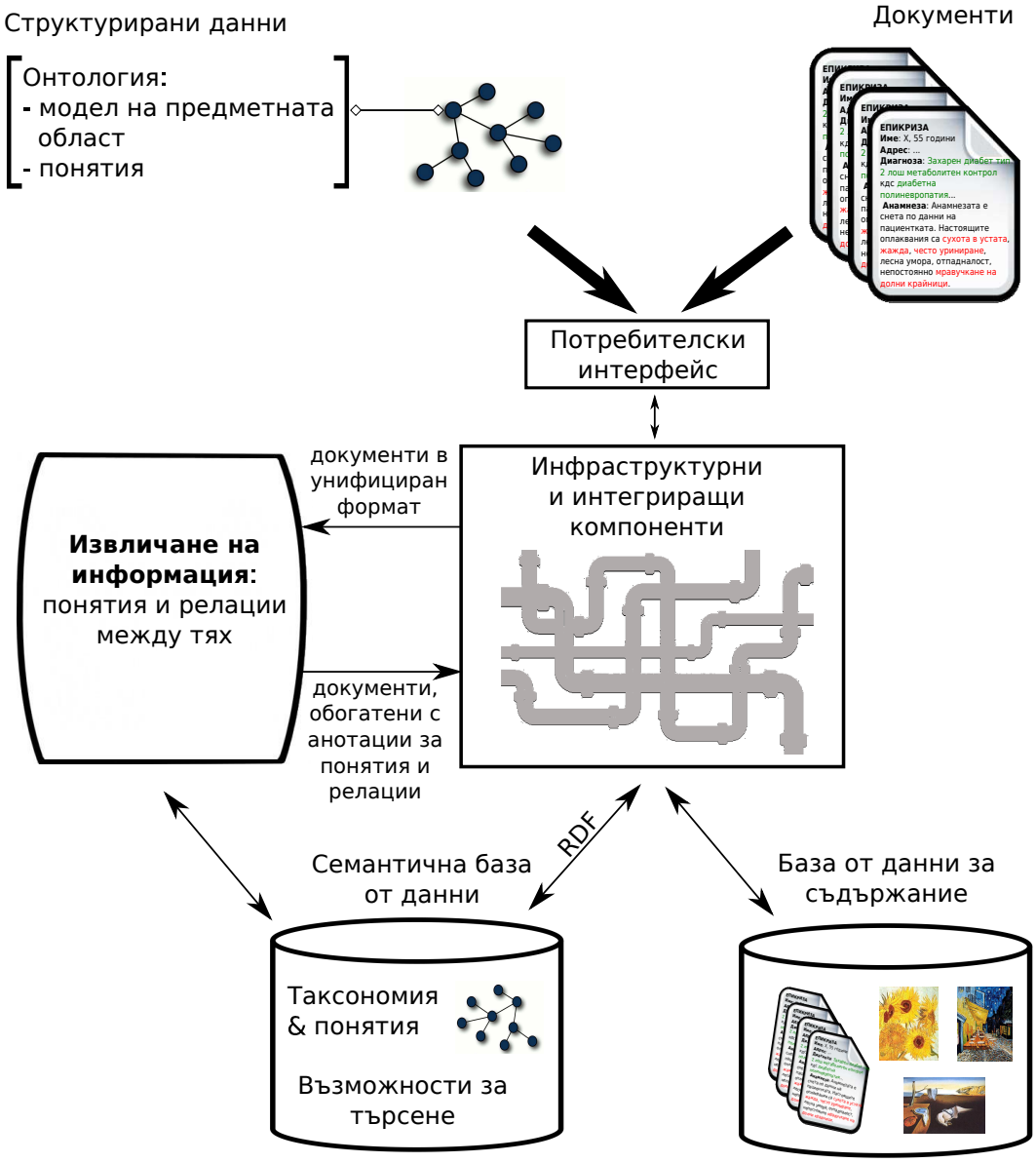
1.1 Характеристики на семантичните системи

За да опишем приложението на техниките за обработка на естествен език в семантичните системи, ще се спрем накратко върху структурата и предназначението на тези системи. На фигура 1.1 са демонстрирани модулите, които обичайно съставят една ССТ - онтологично знание за света; компонент за извличане на информация; семантична база данни; база от данни за съдържание (текстови документи, визуални материали и др.); потребителски интерфейс, инфраструктурни и интегриращи компоненти.

Онтологичното знание за света в ССТ представлява така наречения модел на предметната област. В него са описани основните типове обекти в разглеждания свят, техните характеристики и връзките помежду им, също така и факти за някои от обектите. В системи, работещи с новинарски статии, важните обекти обикновено са хора, географски обекти и организации, докато в биомедицината това са анатомични органи, лекарства, симптоми, гени и др.

Поради голямото разнообразието от области на приложение на ССТ и обвързаността на тези модели с езика, на който функционира системата, често изниква нуждата онтологичното знание да се опише в хода на създаване на самата система. Този процес е трудо- и времеемък и се извършва от висококвалифицирани експерти, които могат да обобщят знанието за областта в онто-

логии или терминологични речници. В тази дисертация ние предлагаме подход за подпомагане на този процес чрез методи на ОЕЕ за автоматично извличане на парафрази и релации между понятия от свободен текст, с които да се обогати модела на ССТ.



Фигура 1.1: Структура на семантична система, обработваща текст.

Предлагаме метод за извличане на йерархичната релация **дете-родител** (**IS-A**) и релацията **засяга** (**AFFECTS**), която означава връзка между заболяване и съответен орган, в който заболяването причинява патологични изменения.

Компонентът за извличане на информация е фокусът на тази дисертация. Тук първо се прави общ анализ на текста за автоматично определяне на лексикалните и морфосинтактичните му характеристики. Наричаме тази фаза предварителна обработка на текста, в нея модулите се изпълняват винаги в една и съща последователност и се счита, че е основата за всеки по-задълбочен текстов анализ. За английски език тези техники са достигнали най-висока успеваемост в сравнение с другите езици. Това се дължи на широката достъпност на ресурси за английски, на разпространеността на литературата на английски език, а също на международното значение на езика, дългогодишното фокусиране на компютърната лингвистика върху английския език, както и на големия брой специалисти, за които това е работен/майчин език.

Същинската обработка и извличането на информация се случват след това, като се използват натрупаните от предварителната обработка характеристики на градивните елементи на текста. Този етап се нарича още разпознаване на именовани единици. За целта се ползват речници с имена на важни обекти, извлечени от бази данни, списъци от изрази, сигнализиращи наличието на именовани единици, а също така и правила, комбиниращи характеристики на низовете в текста. Тези характеристики могат да са повърхностни (дължини на думите, низ, ортографски запис), морфосинтактични признаци (части на реч-та, лема, корен, конституенти и др.), структурни, контекстни (низове отляво и отдясно на входния низ), семантични (семантичен тип) и др. Когато са налични достатъчно количество анотирани данни се прилага и машинно самообучение, базирано на подобни характеристики.

След разпознаването на имената в текста, специфична характеристика на ССТ е свързването на тези имена с точно определени обекти в семантичната база данни или така нареченото семантично анотиране. От важно значение тук е модулът за разрешаване на многозначността на ниво тип на обекта и на ниво обект.

В дисертацията демонстрираме подходи за извличане на именовани единици и частично структуриране на информация от пациентски записи на български

и английски език. При записите на български език се интересуваме от характеристики на пациента като възраст, период на заболяването и симптоми на диабета. От записите на английски език извличаме събития, които имат разнороден вид - прием на пациента, прием на лекарство, оплакване на пациента и други, които ще опишем по-подробно в следващите глави.

Като всяка цялостна система, и ССТ имат **интегриращ компонент**, който се грижи за комуникацията между отделните модули и също така **интерфейс към крайния потребител**. Тези два компонента са извън фокуса на изследванията, представени в дисертацията, затова няма да ги разглеждаме в детайли.

1.2 Разпознаване на парафрази

Определянето на близост между документи и изрази и в частност разпознаването на парафрази е в основата на много задачи в съвременната обработка на естествен език и семантичните технологии. От успешното разрешаване на този проблем зависят качеството на извличането и структурирането на информация от свободен текст, автоматичното отговаряне на въпроси, автоматично извличане на документи, машинния превод, автоматично създаване на онтологии, както и на много други приложения в областта на ОЕЕ. Парафразите могат да бъдат изрази с различна дължина. Това са както синоними на термини, означаващи едно понятие, изразени с една или няколко думи, така и изречения, съдържащи същата информация или заглавия на статии на една и съща тема.

В съвременните изследвания за разпознаване на парафрази се използват както подходи, базирани на правила, така и машинно самообучение и хибридни подходи. В [36] се предлагат мерки за лексикална близост върху потенциални двойки парафрази, за да се определят най-добрите кандидати. Други автори правят концептуален анализ на двата ресурса, в които са кандидатите-парафрази [42]. Те идентифицират понятията в двата текста и прилагат мерки за лексикална близост върху конституентите, които ги съдържат, за да проверят дали се припокриват информационните сегменти, открити в двете фрази. Друг класически подход е изчисляването на близост между парафрази чрез мерки за близост на вектори. В [64] се ползват варианти на $tf - idf$ за представяне на документи, а в [65] е предложен алгоритъм за създаване на канонични форми на изреченията, така че текстове с подобно значение да имат по-голям шанс да бъдат трансформирани в една и съща структура, за разлика от тези, които имат различен смисъл.

Типична задача, включваща разпознаване на парафрази и синоними е откриването и нормализирането на имена на заболявания, симптоми и анатомични органи. Сложността на задачата произтича от няколко фактора, които са характерни за биомедицински текстове. Имената на заболяванията често комбинират в себе си думи от гръцки и латински например “lymphogranulomatosis”. Те могат да бъдат бъдат композирани от категория на болестта и модификатор, който може да е в разнообразни форми, включително анатомичен орган, какъвто е

случаят с “рак на гърдата”, или да е симптом “cat-eye syndrom”, вид лечение “инсулинозависим захарен диабет (тип I)”, агент, който е причинител на заболяването “стрептококова инфекция”, етимология идваща от биомолекулярното изражение на заболяването “идеопатична парапротеинемия”, епонимия “синдром на Сьогрен” и др. Освен това модификатори често могат да се ползват, за да определят състоянието на заболяването, а не и като част от името - например “обострен синузит”.

Друго затруднение за откриване на парафразите са съкращения, инфлексията, ортографския запис, подредбата на думите в имената и синонимите. Всичко това важи както за понятията, така и за отделните думи, а те могат да са части от съставни думи. Това увеличава още повече сложността на откриване на парафразите. Например “oculocerebrorenal” може да се изкаже с няколко думи (око, мозък и бъбрек), включени във фрази с различна подредба.

Задачата за нормализиране на наименования на болести обикновено се решава с хибридни подходи, които включват лексикални и лингвистични методи ([26], [15], [28]). Нормализирането на стринговете, премахването на специфичната ортография, свеждането до корен и други техники за промяна на повърхнинните характеристики обикновено помагат, но вариантите на изписване на имената от лексикона винаги поставят някакви ограничения. За преодоляване на тези ограничения с машинно самообучение има сравнително малко изследвания, като едно от тях е [33].

Успеваемостта на системите за извличане на информация се измерва с два основни критерия - точност P (precision) и покриваемостта R (recall). Точността показва колко от извлечените примери са положителни (такива, които сме искали да извлечем), а покриваемостта - колко от всички налични примери в текста сме успели да извлечем. В различните задачи, балансът между изчерпаемост и коректност може да варира. Например, когато извличаме автоматично информация за състоянието на пациента, за да го съпоставим с лечението на някой друг, искаме да се доближим максимално до резултата, който би произвел човек, ако извлича ръчно тази информация. Тогава ни интересува висока точност. Но ако искаме да подпомогнем експерт или система да филтрира голям обем данни, тогава е достатъчно да предложим по-широк избор от факти, които може да не са съвсем точни, и експертът сам да прецени, кои резултати

са му нужни. В този случай е важна по-висока покриваемост. Като мярка за сравнение между системи най-често се ползва $F-$*мярка* - хармонично средно на точност и покриваемост (Вж. дефиницията в уравнение 1.2.1).

$$F = 2.\frac{P * R}{P + R} \tag{1.2.1}$$

В [30] са публикувани едни от най-добрите резултати за точност при разпознаване на парафрази на заглавия, която е постигната досега. Авторите използват комбинация от лексикални и семантични характеристики на текста и постигат точност от 76.6% и покриваемост - 79.6%. Този подход има по-ниска покриваемост в сравнение с други подобни подходи, напр. в [63] методите работят с 83% покриваемост, но с малко по-ниска точност - 75.6%. Най-общо казано методите, споменати по-горе разрешават проблема за разпознаване на парафрази чрез мерки за лексикална близост, семантично представяне и различни техники за машинно самообучение, които ние също използваме.

1.3 Извличане на понятия и релации между тях

Още от древността хората са се опитвали да систематизират знанието, което имат и да опишат света около себе си. Аристотел също отделя внимание на този проблем в трактата си от Органон - Категории, като поставя обектите от реалния свят в категории, наречени “ta legomena” (нещата са казани) . През Средновековието и Ренесанса други философи и ботаници също се интересуват от този въпрос и предлагат собствени решения. По-късно при осъзнаването на факта, че системите за изкуствен интелект имат нужда от знание за света, се създава необходимост да се извлича и описва знание по начин, който е подходящ за четене от машини/компютри. Тогава става ясно, че при създаването на тези описания е неизбежно съприкосновението с естествения език.

През 1959 Сосюр дава едни от първите класически определения за релации между понятията в [45]. Според него те се делят на синтагматични и асоциативни (парадигматични). Първите са валидни само в рамките на даден контекст, а асоциативните се базират на натрупан опит. По-късно се появяват и други класификации, валидни за конкретни видове задачи и предметни области, но това разделение се запазва.

С течение на времето различни групи работят върху създаването на големи по обем ръчно-направени онтологии, в които да опишат знанието за света или за конкретни предметни области. Такива примери са Cyc, OpenMind Common Sense, MindPixel, FreeBase. През 1962 Килиан въвежда понятието “*семантична мрежа*”[1], като граф, чийто смисъл се моделира от именовани асоциации между думи. Върховете на мрежата са понятия, към които са асоциирани думи (етикети) от текст, а ребрата са връзките между тези понятия. Типичен пример за такава мрежа, която се намира на кръстопътя между моделирането на знанието и езика е WordNet [37], [16]. Тя съдържа 155 000 думи - съществителни, глаголи, прилагателни, наречия, и семантични релации като синонимия, антонимия, хипернимия, меронимия и др.

Увеличаващите се ресурси в електронен формат и желанието за по-бърз напредък в тази област водят до първите опити за автоматично създаване на концептуални модели. В [24] се описва метод за автоматично извличане на лек-

[1] http://en.wikipedia.org/wiki/Semantic_network

сикални релации на хипонимия от неструктуриран текст. Методът цели да избегне предварително кодиране на знание и също така да е приложим за текстови колекции от разнообразни области. Авторите на [7] създават приставка (plug-in) OntoLT за широко-разпространената среда за разработване на онтологии Protégé. OntoLT подпомага интерактивното извличане и/или разширяване на онтологии от текст. Лингвистичният анализ се интегрира с концептуалното описание чрез дефиниране на правила, които съпоставят лингвистични обекти в анотирани текстови колекции с понятия и потенциални атрибути (в Protégé това са класове (*class*) и свойства (*property*)). По този начин е създадена онтология на неврологични понятия, чрез използване на колекции от резюмета на научни статии в неврологията. Фокусирайки се върху извличането на релации, системата RelExt извлича онтологии чрез автоматичното идентифициране на релевантни тройки (двойки от понятия, свързани с релация) от колекция от специфични за областта текстове. RelExt извлича глаголи и техните граматични аргументи (напр. термини) и изчислява съответните релации чрез комбиниране на лингвистична и статистическа обработка [48].

Система с подобно име – RelEx – подпомага извличането на релации от свободен текст [18]. RelEx прави синтактичен анализ и извежда дependентни дървета на входните изречения, като прилага набор от правила върху тези дървета. RelEx е оценена върху множество от 1 млн. резюмета от MEDLINE[2], отнасящи се за релации между гени и протеини и е извлякла около 150 000 релации с точност 80% и 80% покриваемост [18].

За нас е интересна автоматичната обработка на релации в UMLS[3], най-големият ресурс от медицински номенклатури, събирани от Националната библиотека по медицина на САЩ от 1986 г. UMLS е многоезикова терминологична банка, която се използвана за разнообразни приложения - индексиране, архивиране, класифициране и организиране на информация, свързана със здравеопазване и здравен мениджмънт. Съдържа над 150 източника (речника) и над 2 млн. понятия. Ресурсът е организиран около понятията, като множеството от

[2]MEDLINE - Основната онлайн база от библиографични цитати на Националната библиотека за медицина на САЩ, която осигурява международен достъп до списанията на биомедицинска тематинка по света.

[3]UMLS - http://www.nlm.nih.gov/research/umls/

всички термини от всички източници и отделните източници реферират към тези понятия. Статията [13] анализира възможностите за използване на онтологичните релации, за да се произведат автоматично правилни семантични структури за даден медицински документ. Авторите представят метод наречен SeReMeD, който описва начин за генериране на структурирано представяне на неструктурирани медицински записи. Този метод използва релациите между понятия в UMLS и мрежата от семантични типове на UMLS, за да разпознае допълнителни семантични релации, които да подпомогнат процеса на структуриране. Резултатите показват, че релациите могат да подобрят автоматичното генериране на семантични структури. Има много изследвания върху извличане на релации, но ние се фокусираме върху тези от кратки медицински дефиниции и описването на ограничения (правила за филтриране), които могат да помогнат да се уточнят и обяснят откритите релации. Затова ние проучихме и подходите за обработка на релации. Статията [60] представя метод за филтриране на релации и метод за откриване на нови релации, които са намерени в контекста на междуезиково извличане на документи (МЕИД). Използват се семантични анотации, по-точно семантични релации в медицинската област. Като база за сравнение се използват семантичните релации между медицински понятия в UMLS. И двата подхода са приложени за корпус от медицински резюмета на английски и немски език и оценени за тяхната ефективност в МЕИД. Резултатите показват, че филтрирането намалява покриваемостта без да увеличи съществено точността, докато откриването на нови релации наистина се оказва успешен метод за подобряване на извличането на документи. Полезни съвети в тази насока има и в [61]. Авторите представят изследване за идентифициране и оценяване на контекстни характеристики и методи за машинно самообучение, за да се разпознаят медицински семантични релации в текст (по-точно в резюмета от Medline). Ползвайки йерархично клъстеризиране авторите сравняват и оценяват лингвистичните аспекти на контекста на релациите и различни представяния на данните. Чрез подбор на характеристики (feature selection) върху малък обем от данни те показват, че релациите са характеризирани от типични контекстни думи и чрез изолиране на тези думи може да се конструира подходящ езиков модел, представящ целевите релации.

Както показахме по-горе, за да се атакува задачата за извличане на релации

от биомедицински текстове, изследователите ползват широко разнообразие от техники, които са свързани с характеристиките на предметната област. Въпреки че някои изследвания се фокусират само върху една техника, повечето интегрират няколко метода за решаване на задачата. Тези експерименти демонстрират значимостта на статистически извлечените правила за текстообработка, но също така и интегрират лингвистичните характеристики и експертното знание; тази комбинация е съществена за разбирането на био-медицински език [9]. Средата OntoLT, която се ползва за извличане на онтологии от текст, предлага език за дефиниране на предварителни условия (правила за съпоставяне). Предварителните условия са реализирани чрез изрази на XPATH върху анотации, базирани на XML. Ако всички условия са удовлетворени, правилата активират един или повече оператори, които описват по кой начин трябва да се разшири онтологията, ако се намери кандидат за текстова единица, означаваща концептуален елемент. Ние намираме тази идея като подходяща и за нашия подход за извличане на релации.

1.4 Автоматично структуриране на описания от пациентски записи

Задачата за структуриране на описания в пациентски записи (ПЗ) е от съществено клинично и социално значение, тъй като резултатите от нея могат да бъдат използвани повторно за подобряване на здравните грижи за пациента, здравния мениджмънт и здравеопазването като цяло. От научна гледна точка това е подобласт на ОЕЕ и е известна като "*извличане на информация*" в биомедицината (biomedical information extraction).

С увеличаване на обема на здравните записи в електронен формат се появява и нуждата те да бъдат ефективно съхранявани и обработвани, включително да има и ефективен начин за търсене на структурирани описания в тях. От друга страна се появява желанието записите да бъдат използвани като: източник за наблюдения, които могат да се правят само над голям обем данни, като справочна информация за минали случаи, взаимодействия на лекарства, последователности от лечения и/или да се ползват за създаване на регистри за конкретен тип заболявания. Тенденцията вече се е разпространила в повечето държави от развития свят, но започва от САЩ, където се влагат и най-много средства за разработването на такива системи и съответно най-успешните алгоритми/системи за идентифициране на пациентски характеристики и здравно състояние работят предимно върху документи на английски език. ПЗ са богат източник на информация относно здравното състояние на пациента и неговото лечение до момента, но често съществуват само във формата на свободен текст, което налага и създаването на алгоритми за автоматична обработка на свободния текст и извличане на информация. Една от целите на тази дисертация е да предложи подходи за извличане на симптоми на диабета от ПЗ на български език и извличане на медицински събития от ПЗ на английски език.

Извличането на информация в биомедицината е гореща област, в която се работи интензивно в няколко основни подобласти - автоматична обработка на: пациентски записи, амбулаторни листове, клинични изследвания, научни статии. Обзорите [67], [66] и [49], специални издания на списания [29] и [9] и книги [1] в областта посочват, че софтуерът за ОЕЕ, който е създаден с общо предназ-

начение, не е подходящ за обработка на биомедицински текстове, защото те са силно специализирани. Системата представена в [46] решава задачата за идентифициране дали пациентите са пушачи или не чрез класификация на отделни изречения от статуса на пациента. Тя постига $F-$*мярка* 85.57%, като едно от ограниченията й е, че не обработва отрицанието. Подобно на този подход, нашите входни документи са сегментирани на изречения, които се класифицират. Описанията на симптомите са кратки и написани в рамките на едно изречение, затова е важно да се филтрират изреченията, които не носят информация. Ние използваме техники за машинно самообучение и анализ, базиран на правила, и проверяваме за наличие на отрицанието.

Harkema et al. [23] представят алгоритъм наречен ConText, който определя дали дадени клинични състояния, които са описани в медицински запис, са употребени с отрицание, са хипотетични, минали, или са изпитани от някой, който не е пациента. Системата е изцяло базирана на правила и определя статуса на дадено състояние по наличието на сравнително прости лексикални срещания, които се намират в текста в близост до споменаване на състоянието. Този алгоритъм се прилага успешно за различни типове клинични доклади с F-мярка съответно за: разпознаване на отрицание (75-95%), минали събития (22-84%), хипотетични събития (86-96%) и събития, отнасящи се за пациента (100%) в зависимост от типа на документа. Нашият подход е основан на подобна идея – ние подготвяме няколко научени от данните речника и ги използваме за определяне на границите на изразите, които искаме да извлечем, но за разлика от [23], се фокусираме върху извличането на описанието на симптоми, техния статус, стойности и евентуално отрицание.

Отрицанието е една от най-важните характеристики за разпознаване в медицинския текст. Boytcheva [5] третира този проблем чрез анализ на сигнализиращи изрази. Същият подход прилагаме и ние.

Много системи съдържат модули за извличане на отделни състояния/симптоми на пациента, но рядко интегрират завършен модел на всички състояния. Например MedLEE [17] и MEDSYNDIKATE [21] идентифицират статуса на състоянието и модифициращата информация като анотомично място, отрицание, промяна през времето. В Boytcheva et al. [4] авторите извличат от български ПЗ статуса на кожата, крайниците, шията и щитовидната жлеза с висока точност.

Глава 2

Приложение на OEE за създаване на концептуални модели на предметна област

2.1 Въведение

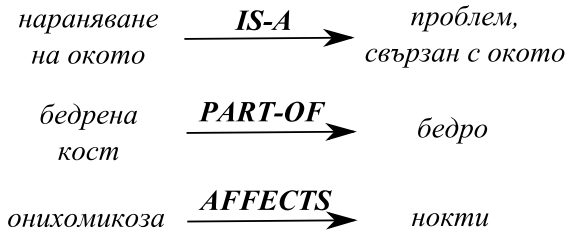
За да се направи задълбочен автоматичен анализ на текстови документи на определена тематика, е необходимо наличието на концептуален модел, описващ предметната област, който да подпомогне интерпретирането на обектите, идентифицирани в текста и връзките помежду им. Под модел в този случай разбираме концептуални структури, записани в декларативен формализъм, които агрегират основните понятия от предметната област и релациите между тях. Такива са онтологиите и базите от знание. Те помагат при намирането на важни късове информация в текста и структурирането им и са от ключово значение за разработването на надеждни системи за извличане на информация, които да работят с висока точност.

В повечето от изследванията, представени в тази дисертация, се анализират пациентски записи (ПЗ) от болничен престой на хора, диагностицирани с диабет. Предметната област в случая обхваща медицинска терминология, описваща заболяването диабет: симптоми; свързани заболявания; анатомични органи, които са засегнати; лекарства и процедури за лечение; изходи от лечението; наименования на специализирани болнични звена, където се осъществява ди-

агностиката и лечението; наименования на специализирания персонал, който се грижи за диагностициране и лечение.

Освен термините, в модела се описват и връзките между отделните обекти от областта. Връзките, които имат отношение към обработката на естествен език и разбирането на текста в задачите, представени в тази дисертация са йерархичната връзка **дете-родител** или още известна като **IS-A**, която свързва по-специфични с по-общи понятия, и връзката **засяга/влияе на**, известна като **AFFECTS** и свързва заболявания или симптоми с органите, в които те предизвикват патологични изменения.

Примери, демонстриращи някои от релациите са представени на Фигура 2.1.

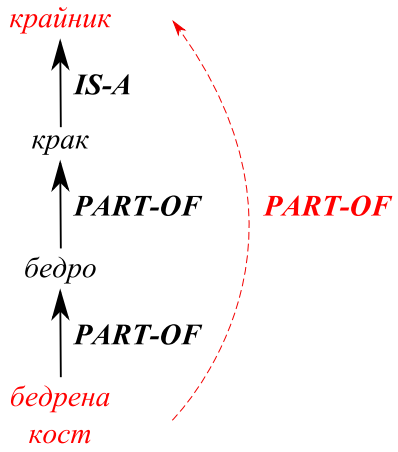


Фигура 2.1: Примери, демонстриращи семантични релации.

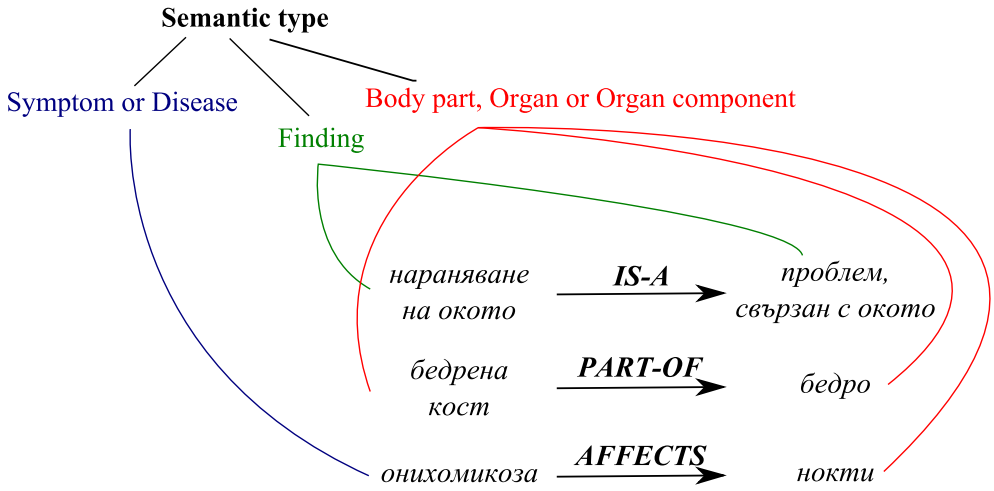
Наличието на тези връзки между понятията ни позволява да правим обобщения и спецификации върху фактите, които извличаме от документите. Така, ако разполагаме с факти за усложнения в бедрената кост, ние можем автоматично да заключим, че пациентът има усложнения в крайниците без думата крайници да е употребена в текста, защото от нашия модел знаем, че бедрената кост е част от бедрото, а бедрото е част от крак, който е крайник (вж. Фигура 2.2).

Обогатявайки модела със семантични типове на понятията от областта ние разширяваме възможностите за анализ върху текста и за разбиране на фактите в него (вж. Фигура 2.3). Така например ако в описанието на статуса на пациента срещнем текста "*..., онихомикоза*", използвайки модела на областта, можем да заключим, че онихомикозата е заболяване и то засяга ноктите, които са част от тялото. Ако в следващото изречение се говори за "гъбички", с голяма вероятност системата може да заключи, че става дума за ноктите.

Когато моделът на областта е богат на ресурси, ясно построен, и отразяващ адекватно спецификата на обектите и връзките между тях, той позволява



Фигура 2.2: Пример, демонстриращ обобщение.



Фигура 2.3: Извадка от модел на предметна област.

по-задълбочен анализ на текста. Моделът, описващ нашата предметна област, трябва да подпомага задачата за извличане на информация, както и последващо търсене в концептуални шаблони. Той трябва да предоставя възможност за намиране на по-общи, по-специфични понятия или понятия-сестри и по този начин да служи при установяване на близост между структурирани пациентски записи.

Понятията от изследваната тук предметната област са налични в терминологични речници на български и други езици (напр. анатомични речници), йерархии (международна класификация на болестите МКБ/ICD 9 и 10[1], SNOMED

[1]ICD http://www.who.int/classifications/icd/en/

CT[2] [52], MeSH[3]), учебни материали (речници и йерархии, изградени с учебна цел), вътрешно-болнични класификации. Всички тези ресурси са създавани за задачи, различни от обработката на естествен език. Например SNOMED CT е създадена с цел да стандартизира медицинската терминология и да допринесе за подобряване качеството и сигурността в здравеопазването, като предостави рамка за консистентно представяне на значенията в електронните здравни записи в САЩ. MeSH е създадена с цел индексиране на научни публикации за базите от данни на MEDLINE/PubMED® и създаване на каталози на книги, документи и аудиовизуална медия, придобити от Националната библиотека по медицина(NLM[4]), САЩ. Тези ресурси не могат да бъдат директно приложени за решаване на задачата за семантичен анализ на текст, но могат да бъдат източник на термини и релации за обогатяване на модела на предметната област. Тъй като не съществува единен ресурс на български език, описващ пациентските записи на диабетици, се налага да създадем модела на нашата предметна област чрез обединяване на информация от различни източници и допълнителна обработка. Създаването на модел може да стане ръчно, чрез комбиниране на речници, но това е трудоемка задача, поради което се налага да намерим полуавтоматични начини за извличането му от множество източници и работния корпус от документи.

От изброените горе ресурси, в електронен формат на български са налични само класификациите ICD версия 9 и 10, които сами по себе си не са достатъчни, за да се направи задълбочен автоматичен анализ на текста, към какъвто се стремим. Затова използваме подход **отдолу-нагоре** (bottom-up) за създаване на модел, при който първо извличаме автоматично речник на понятията от наличните документи и после ги обогатяваме със синоними, свързани термини и връзки между тях. При обработка на документи в тясна предметна област, както в случая - епикризи на диабетици от една и съща болница, този подход се изразява в статистически анализ на текста, с който се цели да се състави честотен речник на употребените в него думи и фрази. Честотният речник е основата на речника, описващ терминологията в колекцията от документи.

[2]SNOMED CT http://www.ihtsdo.org/snomed-ct/
[3]MeSH https://www.nlm.nih.gov/mesh/
[4]NLM https://www.nlm.nih.gov/

В тази дисертация са разгледани два подхода за приложение на OEE за обогатяване на модела на предметна област чрез анализ на епикризи на диабетици. Първият подход е свързан с разпознаването на синоними и парафрази на понятията от наличната терминологичната банка, а вторият е свързан с обогатяване на модела с нови релации, свързващи понятията от областта - релации от типа **дете–родител** (**IS–A**) и **засяга** (**AFFECTS**).

2.2 Разпознаване на синоними и парафрази на понятията от наличната терминологична банка

2.2.1 Необходимост и приложение на разпознаването на парафрази

При анализ на текст един от основните проблеми е разпознаването на парафрази и синоними и съответно идентифицирането на еднакви или близки събития (например *онихомикоза* и *гъбички по ноктите* могат да се интерпретират като едно и също понятие). Богатството на езика позволява да се изразяваме с разнообразни лексикални и граматически форми и това затруднява автоматичната му обработка. В по-тесни области като на диабет, лексикалното разнообразие е сравнително по-малко. И все пак изказът на лекарите се различава поради: наложените болнични разпоредби, ползването на предефинирани понятия в болничния софтуер или заради лични субективни предпочитания. Един модул за автоматично извличане на информация от епикризи има възможно най-голямо покритие, когато неговият модел, описващ предметната област, съдържа понятията от работния корпус с документи и техни синоними, парафрази и свързани понятия. Тук предлагаме подход за създаване на такъв модел чрез полуавтоматично извличане на синоними и парафрази на термините, които се срещат в работната колекция от документи чрез използване на външни ресурси на английски език.

Разпознаването на парафрази е добре известен проблем, който се решава както в едноезиков, така и в многоезиков контекст. Задачи, които изискват разпознаване на парафразите, са например: свързване на онтологии, извличане на документи, автоматично отговаряне на въпроси, опростяване на текст, построяване на терминологични речници. Липсата на пълна информация по време на обработката и лексикалното разнообразие, с което може да се изрази смисълът на дадени понятия, правят сложността на тази задача доста висока дори и за хора. Това може да се види като се измери κ-статистика на съгласуваност между експертите, които ръчно анотират даден корпус. Например Microsoft Research

Paraphrase Corpus [14] е корпус от изречения-парафрази, който съдържа ръчна оценка на всяка двойка изречения и има изчислена κ-статистика от едва 83% съгласуваност между експертите, поставили оценките.

Съвременните изследвания за намиране на парафрази варират от търсене на единични думи, синоними на термини, до идентифициране на цели изречения, които са парафраза едно на друго, като например заглавия на новинарски статии. За сравнение, нашата задача е да намерим термини на английски в UMLS, които могат да са етикети на термини, използвани в нашия корпус от епикризи.

2.2.2 Дефиниция на задачата

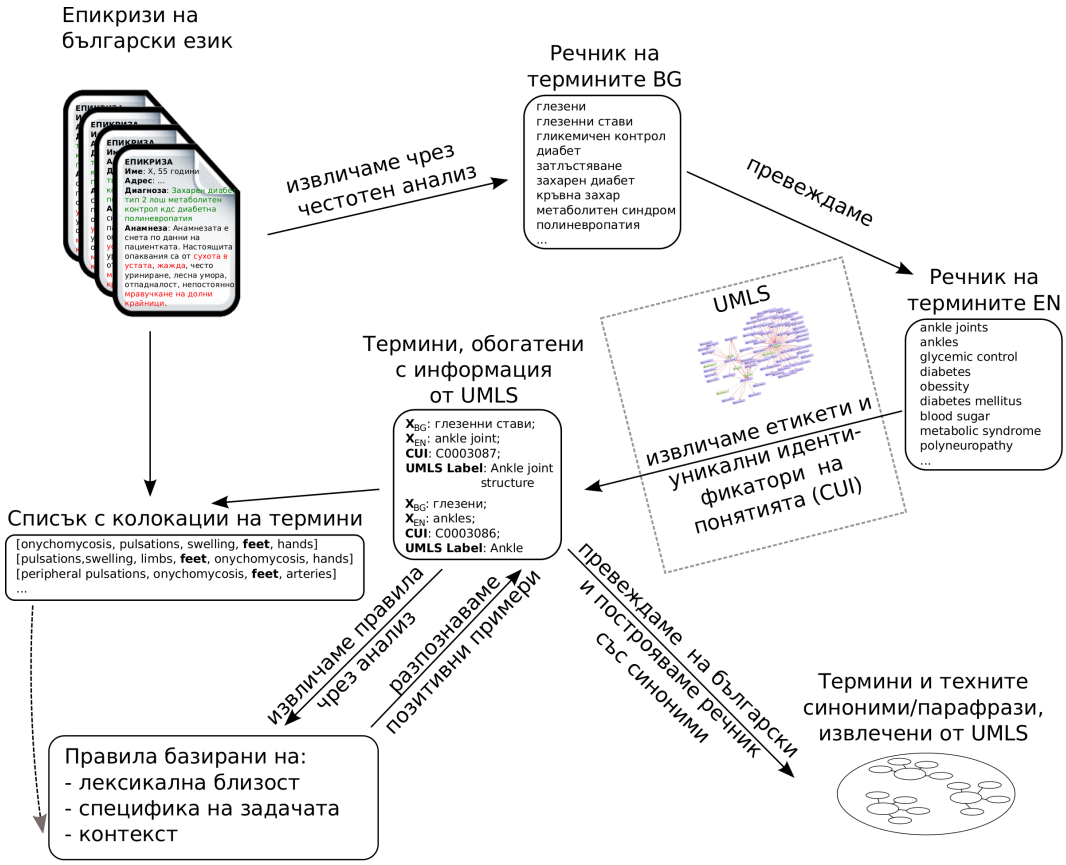
Задачата, която си поставяме, е да построим терминологичен речник от понятията, които се срещат в колекция от 1000 пациентски записа на диабетици. След това полуавтоматично да го обогатим с парафрази/синоними на термините в него. Така създаденият ресурс е основата на модела на предметната област. За осъществяването на тази задача демонстрираме подход за извличане на термини на английски от UMLS, които могат да са етикети на термини, използвани в нашия корпус от ПЗ. Процесът на извличане е демонстриран на Фигура 2.4.

2.2.3 Метод

Извличане на терминологичен речник

За построяване на концептуален ресурс на български използваме подход **отдолу-нагоре**. Започваме с предварителна обработка на корпуса от епикризи и с набор от статистически методи, техники за ОЕЕ и ръчно отсяване на значимите понятия извличаме речник на термините, употребени в корпуса. Пример с най-често използваните думи в корпуса е показан на Фигура 2.5.

Честотният анализ на корпус цели да съпостави на всяка дума теглото ѝ в колекцията от документи. За целта свеждаме текста до **торба от думи** (bag-of-words), където на всяка дума съпоставяме брой на срещане в текста. Отстраняваме най-често срещаните думи (стоп-думите) и след това прилагаме мярка за определяне на теглото на думите. В случая е подходящо да ползваме броя



Фигура 2.4: Извличане на парафрази.

на срещанията на думата в колекцията като такава мярка. Тя е още известна като tf (total frequency).

Теглата помагат да разграничим важни от маловажни думи. Термините от областта обикновено попадат между думите с високи стойности на tf и са близо един до друг. Определя се праг на значимост на думите по tf и всички думи, чието тегло преминава прага, се считат за потенциални термини. След това експерт от областта ги преглежда ръчно и отсява само термините. Към тях се добавят и фразите, които съдържат тези термини. Така полученият списък от понятия е основата на модела, който описва разглежданата предметна област. Следват стъпки по обогатяването му с нови понятия и етикети, описващи връзката между тях.

Дума	Честота на срещане в корпуса	Причина за игнориране
на	14476	твърде често срещана
с	10428	твърде често срещана
и	9175	твърде често срещана
в	5930	твърде често срещана
е	5568	твърде често срещана
за	5436	твърде често срещана
се	4602	твърде често срещана
ч	4088	твърде често срещана
от	3760	твърде често срещана
по	2637	твърде често срещана
лечение	2524	термин
г	2225	твърде къса дума
диабетна	2038	термин
диабет	1992	термин
при	1949	твърде къса дума
не	1923	твърде къса дума
т	1719	твърде къса дума
тип	1696	твърде къса дума
без	1632	твърде къса дума
със	1601	твърде къса дума
х	1600	твърде къса дума
захарен	1590	термин
л	1586	твърде къса дума
мг	1488	твърде къса дума
състояние	1449	термин
полиневропатия	1327	термин
след	1225	термин
данни	1183	термин
гр	1181	твърде къса дума
анамнеза	1137	термин
клиника	1072	термин
захар	1069	термин
кош	1040	термин
контрол	1019	термин
тегло	1012	термин

Фигура 2.5: Извличане на термини от колекция с документи, базирано на честота на срещане на думите в текста.

Извличане на синоними и парафрази на термини

За намиране на правилните съответствия (парафрази/синоними) между термини от новосъздадения речник и термини на английски от UMLS прилагаме

правила, създадени след анализ на входния корпус. Те включват характеристики за лексикална близост, семантична близост, базирана на колокации на термините в контекста, характеристики на представянето на термините в UMLS и органичения (стоп-думи).

От UMLS използваме частично следните ресурси: *понятия*, *синонимни множества*, *семантични типове* и *релации*. Има няколко причини, които възпрепятстват директното заемане на подмножество от термини от UMLS:

- входната терминология и корпуса от ПЗ са на български език, а ресурсите от UMLS, които ползваме, са на английски език;
- поради наличието на няколко лексикални записа на едно понятие в корпуса и в UMLS, не може еднозначно и лесно да се определи кои термини от UMLS бихме искали да извлечем;
- бихме искали да извлечем от UMLS синоними на термините от нашия корпус, а не само техните преводни съответствия.

Поради тези причини има нужда да се приложат допълнителни процедури, за да се идентифицират парафразите на термините от нашия начален речник в UMLS.

2.2.4 Ресурси

Описание на епикризите

Дължината на епикризите в българските болници е обикновено 2-3 страници. Записът е организиран в следните секции: (*i*) лични данни; (*ii*) диагнози на главното и придружаващите заболявания; (*iii*) анамнеза (история на заболяването), включително текущи оплаквания, минали болести, фамилна анамнеза, алергии, рискове; (*iv*) обективно състояние на пациента, включително резултати от прегледа; (*v*) лабораторни тестове и други наблюдения; (*vi*) коментари на други медицински лица; (*vii*) дискусия; (*viii*) лечение; (*ix*) препоръки. Невинаги присъстват всички секции. Епикризите са анонимизирани. Секциите са сравнително ясно отделени, което улеснява търсенето в отделни фрагменти на епикризата. За изработване на модел на предметната област в текущата задача използваме само статуса на пациента.

Е П И К Р И З А

Име: X, 55 години
Адрес: ...

Диагноза: Захарен диабет тип 2 лош метаболитен контрол кдс диабетна полиневропатия...

Анамнеза: Анамнезата е снета по данни на пациентката. Настоящите оплаквания са от сухота в устата, жажда, често уриниране, лесна умора, отпадналост, непостоянно мравучкане на долни крайници.

Статус: Жена в задовлително общо състояние, заемаща активно положение в леглото, ориентирана за време, място и собствена личност. Кожа - без активни стрии, женски тип окосмяване...

Параклинични изследвания: суе 32 мм hb 136 g/l; er 4,7; hct 0,41; leu 7,8; tr 353...
ехо на щитовидна жлеза: десен лоб -55/15/18 мм
ляв лоб - 50/14/15 мм истмус 4 мм ...

Обсъждане: Отнася се за болна със захарен диабет тип 2, с лош гликемичен контрол / hba1c 9,86 % /, на фона на лечение с метформин 3х1000 мг и глимепирид 2 мг, приемани 4 месец. Предвид лошият резултат от кзп на фона на комбинираното лечение от суп и метформин повече от 4 месеца,

Препоръки: Препоръчваме спазване на подходящ хранителен и дивагетелн режим. Да продължи лечението с глимепирид 2мг, сиофор 500+1000+ 1000мг...

Фигура 2.6: Примерна епикриза.

Списък с термини, описващи крайниците

За този анализ разглеждаме само терминология, свързана с описание на статуса на крайниците в епикризите. Построяваме речник по метод отдолу-нагоре: анализира се автоматично статуса на пациента в 1000 епикризи, диагностицирани с диабет - само 828 от тях съдържат описание на крайниците и само 368 от описанията са уникални. Идентифицират се границите на описанията на органите, както е представено в [4]). Извличат се от тях само тези, които се отнасят за крайниците (както на Пример 2.2.1) и се обработват с програма за разпознаване на корена на думата. Изработва се списък от корените на думите от описанията на крайници и този списък се пресича с корените от речника от термини, създаден чрез честотен анализ на изходния корпус. В резултат се получава списък на термините, описващи състоянието на крайниците, който ще наричаме $LTBG1$. Списъкът $LTBG1$ превеждаме ръчно на английски (където е възможно повече от едно преводно съответствие, то е дадено) и означаваме с $LTEN2$. В случая ръчният превод е възможен, защото обема на данните е сравнително малък.

Пример 2.2.1. *Крайници - без отоци, запазени периферни пулсации. Затруднена и болезнена походка, използва помощни средства.*

Списък с термини-колокации от пациентски записи

На следващата стъпка се генерира още една колекция от термини. Тя съдържа множества от термини-колокации за всеки един от термините в гореописания речник $LTBG1$. За всяко описание на крайници и всеки пациент извличаме множеството от термини, които са били употребени заедно в едно описание. От тази колекция създаваме списък, в който срещу всеки термин стоят множествата от термини-колокации, с които той е употребен в отделните описания (ще наричаме този списък LTC). В него се ползват преведените на английски термини и той се поддържа на английски език. Тези множества са необходими на по-късен етап, за да се установи близостта между потенциалните парафрази в Секция 2.2.5. Например в Таблица 2.1 са показани множествата от термини-колокации, които се срещат заедно с термина feet. Всеки ред представя отделно описание на крайниците или контекста, в който терминът се среща в отделните описания.

No на описание	**Списък от термини-колокации**
1	[onychomycosis, pulsations, swelling, **feet**, hands]
2	[pulsations,swelling, limbs, **feet**, onychomycosis, hands]
3	[peripheral pulsations, onychomycosis, **feet**, arteries]

Таблица 2.1: Множество от термини, които се срещат с feet в ПЗ.

UMLS

UMLS предоставя софтуер за локално разглеждане на терминологичната банка, уеббазиран интерфейс, както и API за отдалечено свързване и заявки към сървър UMLS Terminology Services (UTS).

Предимството на UMLS Metathesaurus като терминологична банка е, че съдържа добре документирана, последователно структурирана информация за ресурса. Всички речници са представени в текстов формат RRF, което ги прави лесни за разглеждане и обработка. В допълнение към това има потребител-

ски интерфейс, който подпомага разглеждането на термините и свързаната с тях информация. Ние използваме Metathesaurus за 2 цели: (*i*) да филтрираме по-внимателно термините в корпуса от ПЗ, които искаме да обработваме, и (*ii*) да подберем етикети на понятия-синоними за всеки от тях. За извличането на предложения от UMLS ползваме алгоритъма за точно съвпадение в UMLS Terminology Services (UTS) API [UTS]. При него, в отговор на заявка за търсене по термин, се връщат етикети на понятия, в които има лексикално съвпадение с търсения низ или чиято лема съвпада с търсения стринг. Прави се проверка за съвпадение на лексикално ниво и ниво лема и в *предпочитания етикет*[5] на понятието.

2.2.5 Моделиране на фразова структура и правила за разпознаване на парафрази

Вътрешно представяне

За всяко понятие от списъка с термини $LTEN2$, който сме построили от корпуса с документи, изпращаме заявка за търсене в UMLS. Термините на този етап не са лематизирани, а са подадени като заявка към UMLS в техния лексикален вариант. Целта е да се запази оригиналната форма на термина и да се извлекат повече потенциални кандидати от UMLS. По този начин UMLS ще потърси първо термини, съответстващи на заявената лексикална форма, а после ще лематизира входа и ще потърси съответствие на лемата. При намерено съвпадение (термин, който е същият като търсения или такъв, чиято лема съвпада с лемата на търсения термин) UMLS връща списък от съвпадащи термини (етикети на понятия) и техните характеристики – дефиниция, свързани термини, източници на информацията.

Събраната информация структурираме както е показано на Таблица 2.2:

BG термин (X_{bg}) е термин от входния речник на български език $LTBG1$, *EN превод* (X_{en}) е неговият превод на английски (от експерт) от списъка $LTEN2$. Когато този термин е подаден на UMLS, сървърът връща етикет на понятието,

[5]Предпочитан етикет - от всички етикети за едно понятие винаги има само един предпочитан в даден ресурс. Този етикет се счита за най-подходящо/правилно описващ даденото понятие и обикновено се третира с приоритет при сваляне на многозначността на ниво етикет.

BG термин X_{bg}	**EN превод** X_{en}	**Идентификатор в UMLS (CUI)** X_{cui}	**UMLS термин** X_{umls}
глезенни стави	ankle joint	C0003087	Ankle joint structure
глезенни стави	ankle joint	C1279146	Entire ankle joint
глезени	ankles	C0003086	Ankle

Таблица 2.2: Вътрешно представяне на характеристиките на термините като наредени четворки.

който в таблицата е означен с *UMLS термин* (X_{umls}), а *CUI* (X_{cui}) е идентификаторът на този термин в UMLS.

След анализиране на така структурираните данни, ние изведохме няколко правила, които комбинират различни характеристики като лексикална близост, ограничения, наложени от типа на задачата и контекстни характеристики. Тези правила описват позитивни и негативни примери. Правилата се прилагат последователно към редовете от Таблица 2.2 и само позитивните примери се запазват. Те съдържат термини от UMLS, които са синоними/парафрази на термините от изходния корпус.

Правила, базирани на лексикална близост

Първият тип правила са базирани на лексикална близост. Това са правила 2.2.1, 2.2.3 и 2.2.2.

Правило 2.2.1. *Ако X_{umls} съвпада с X_{en}, тогава примерът е ПОЗИТИВЕН.*

Правило 2.2.2. *Ако X_{umls} е лексикален вариант на X_{en}, тогава примерът е ПОЗИТИВЕН.*

Правило 2.2.3. *Ако X_{umls} съвпада с X_{en} + "(<someStringInBrackets>)", тогава примерът е ПОЗИТИВЕН.*

Пример 2.2.2. *Четворката [глезени; ankles; C0003086; Ankle] е маркирана като ПОЗИТИВНА според правило Правило 2.2.2, понеже ankles е лексикален вариант на Ankle.*

След прилагане на правилата, ако четворката за даден термин се маркира като ПОЗИТИВНА, се проверяват останалите UMLS предложения за същия термин. Етикетите, които имат същия CUI идентификатор като термина от позитивната четворка, също се добавят като ПОЗИТИВНИ. Така обогатяваме броя на етикетите на едно понятие.

Правила, специфични за задачата

Тъй като събираме термини от описания на крайници на пациенти, диагностицирани с диабет, възможността някой термин да е име на ген е пренебрежимо ниска. От друга страна, гените имат разнообразни и неочаквани имена и може да се случи така, че някой ген да съвпадне с име на термин, който ние наблюдаваме. Затова игнорираме предложенията от UMLS, които съдържат думата "gene" (Правило 2.2.4). Имайки предвид това условие, във всички случаи, когато UMLS предлага само един етикет и този етикет не е име на ген, ние го приемаме за позитивен (Правило 2.2.5).

Правило 2.2.4. *Ако X_{umls} съдържа думата "gene", примерът е НЕГАТИВЕН.*

Правило 2.2.5. *Ако има едно единствено X_{umls} и Правило 2.2.4 не се прилага, тогава примерът е ПОЗИТИВЕН.*

Пример 2.2.3. *За преведения термин scars има само едно предложение от UMLS, което формира четворката [цикатрикси, scars, C2004491, Cicatrix]. В този случай четворката с CUI C2004491 е маркирана като ПОЗИТИВНА от Правило 2.2.5, докато в случая [цикатрикс, scar, C1419736, RPS4X gene] терминът "RPS4X gene" води до маркиране на четворката като НЕГАТИВНА от Правило 2.2.4.*

В ресурсите на UMLS често се разграничават следните 2 термина за анатомичните органи – единият е *Entire organ*, съответстващ на самия орган, а другият - *Structure of organ*, съответстващ на структурата на органа. UMLS предлага и двата термина, когато получи име на орган като заявка. За нашите цели само термини от типа *Entire organ* са от значение, тъй като те са семантично по-близки до понятията, които наблюдаваме в анализираните документи (Правила 2.2.6 и 2.2.7).

Правило 2.2.6. *X_{umls} equals "Entire" X_{en}, примерът е ПОЗИТИВЕН.*

Правило 2.2.7. *X_{umls} equals "Structure of X_{en}" or "X_{en} structure", примерът е НЕГАТИВЕН*

Пример 2.2.4. *Преведеният термин "right foot" е получил две предложения от UMLS*
[дясно ходило, right foot, C1279998, Entire right foot] и
[дясно ходило, right foot, C0230460, Structure of right foot].
В този случай само първата четворка, съдържаща "Entire right foot" се приема за ПОЗИТИВНА по Правило 2.2.6.

Правила, базирани на контекст

За да подобрим резултатите от тази процедура, добавяме още едно правило, което взема предвид контекста на термина в документите. То ползва списъците от термини-колокации. Както казахме по-рано, ние събрахме списъци от термини, които се употребяват заедно в едно и също описание. Ако правилата от 2.2.1 до 2.2.7 не могат да определят дали даден пример е позитивен, или негативен, тогава се прилага Правило 2.2.8.

Правило 2.2.8.
*(*i*) IF X_{umls} съдържа точно съвпадение (exact match[6]) на X_{en} ,*
OR lemmatised(X_{umls}) съдържа точно съвпадение на lemmatised(X_{en})
X'_{umls} = {X_{umls}} - {X_{en}} отиди в (ii);
ELSE $X'_{umls} = X_{umls}$ отиди в (ii);
*(*ii*) потърси съвпадение до ниво лема между термините в X'_{umls}*
и в съответните на X_{en} множества от колокации и
IF има съвпадение, примерът е позитивен ПОЗИТИВЕН.
ELSE примерът е НЕГАТИВЕН.

[6]Под "exact match" имаме предвид параметър на алгоритъма за близост на стрингове в UMLS, който приема 2 стринга за съвпадащи, ако те съвпадат напълно или техните леми съвпадат напълно.

Пример 2.2.5. *UMLS предлага следните термини за plantar surface [плантарна повърхност, plantar surface, C0818078, Plantar surface] [плантарна повърхност, plantar surface, C0230463, Sole of Foot]. Първият термин "Plantar surface" ще бъде маркиран като позитивен от Правило 2.2.1, а "Sole of Foot" ще бъде маркиран като позитивен от Правило 2.2.8, защото алгоритъмът е намерил термини в списъка с термини-колокации, които съвпадат с предложените от UMLS термини. Списък с колокации [plantar surface: [dry, plantar surface, feet]].*

Тези правила са само частично зависими от решаваната задача и не са зависими от езика на входните документи, затова биха били лесно приложими за извличане на парафрази на термини в друга подобласт на биомедицинските изследвания. Правилата ни помагат да извлечем парафрази от UMLS, което е най-богатият известен ресурс на медицинска терминология.

2.2.6 Експерименти и резултати

За този експеримент са използвани 368 описания на крайници от 1000 пациентски записа. От тях е компилиран списък с 265 термина, включително лексикални форми ($LTBG1$). Всички тези термини са преведени на английски и са подадени като заявки за търсене в UMLS с параметър "exact match" и в резултат са получени само 186 съвпадения (понякога повече от един отговор за една заявка). Термини, които не са получили нито едно предложение от UMLS, са пренебрегнати. За тестване отделихме 66 UMLS отговора и проверихме правилата върху тях. Ако даден термин влезе в този списък, всички предложения от UMLS за него също влизат в списъка. Резултатите от филтрирането на предложенията от UMLS са представени на таблица 2.3.

Правило No	1	2	3	4	5	6	7	8
Дял на разпознати парафрази (%)	54	2	9	2	11	15	9	2

Таблица 2.3: Резултати от прилагането на правила 2.2.1–2.2.8 върху тестовите данни.

Релевантността на етикетите от UMLS спрямо термините от $LTEN2$, за

тестовото множество от термини, е ръчно оценена. Оказа се, че от 66 UMLS етикета само 46 наистина съответстват на входните термини. От тях процедурата е разпознала и маркирала 43; 2 са били грешно класифицирани и 5 са били пропуснати. Тези резултати са представени на Таблица 2.4.

Релевантни етикети	46
Разпознати	43
Неразпознати	5
Грешно разпознати	2
Точност	96%
Покриваемост	89%
F-мярка	90%

Таблица 2.4: Точност и покриваемост на правила 1–8.

Резултатите са окуражаващи - те показват, че в затворена предметна област със сравнително ясна терминология, подход, базиран на правила и метрики за лексикална близост, може да доведе до сравнително високи резултати. Правила 2.2.1, 2.2.5 и 2.2.6 са най-често прилагани. Това са правилата, които са тясно свързани с представянето на записите в UMLS. Въпреки че Metathesaurus претърпява промени във времето и се поддържа и обновява от различни експерти, нашето изследване показва, че неговата структура е стабилна и той е подходящ ресурс за повторно използване в задача за търсене на термини и парафрази.

2.2.7 Резюме

Представихме подход за намиране на парафрази в междуезиков контекст. Извличаме парафрази на термини, описващи статуса на крайници в епикризи на диабетици на български език. Областта е ограничена и терминологията е доста добре дефинирана, което помага сравнително лесно да се построят правила за филтриране на позитивни от негативни примери. Въпреки малкия брой правила резултатите от процедурата за разпознаване на парафрази са високи. Това е окуражаващо за обработката на биомедицински текстове на български език и е белег, че можем да се опрем на съществуващи ресурси, а няма да започнем да построяваме всички ресурси отначало. Тъй като множеството от

термини в този анализ е сравнително малко, те бяха преведени на английски от експерт, а не автоматично, и това вероятно влияе на резултатите от процедурата. В бъдеще ще проверим какво влияние би оказал автоматичният превод върху тези резулати. Възнамеряваме да направим по-широкообхватно извличане на парафрази от UMLS, като използваме не само точно съвпадение (exact match), но и приблизително съвпадение и ще приложим метрики за близост на низове, за да подобрим разпознаването.

2.3 Разпознаване на релации между медицински понятия чрез обработка на дефиниции от UMLS

2.3.1 Въведение

В предходната глава (2.2) показахме техники за разширяване на терминологичната банка, която описва пациентски записи на диабетно болни. Следващата задача при построяване на модела на проблемната област е добавянето на етикети на връзките между понятията, които термините описват. Някои от тези връзки са налични в съществуващи терминологични ресурси в UMLS като SNOMED CT и MeSH и могат да бъдат извлечени оттам, но те не са изчерпателни. Дори йерархичната релация **IS-A** невинаги е означена в тях.

Много от релациите в съществуващите биомедицински тезауруси, включително и UMLS, изразяват или връзки, които са твърде общи, или такива, които са трудни за интерпретиране в контекста на OEE [40], като например тези за по-специфично понятие (**NARROWER TERM**) и за по-общо понятие (**BROADER TERM**). По този начин автоматичното извличане на релации се превръща в "горещо поле" за научни разработки, особено в такива големи области, където ръчното изброяване е невъзможно.

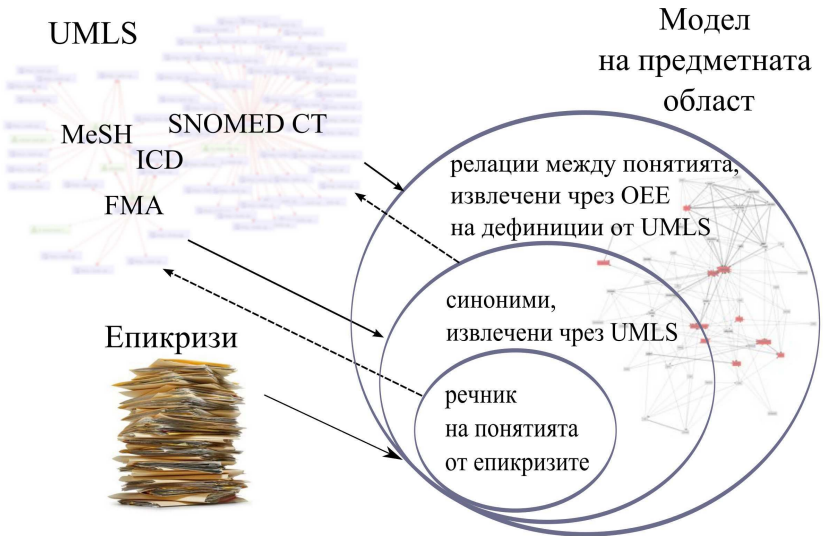
Дефинирането на релации между понятия е сравнително труден процес, тъй като те отразяват връзки, които могат да са неявни и често зависещи от задачата. Има разнообразни решения, които са създадени за конкретни приложения в определени области, но всички досегашни опити да се унифицират подходите за извличане на релации в изкуствения интелект (ИИ) са се проваляли. Все още няма практически работещи решения, където релациите между понятията да се назоват в глаголи и етикетите на понятията да се интерпретират като попълващи ролевите позиции на глагола, напр. подлог-допълнение. В същото време използването на тематични роли (АГЕНТ, ПАЦИЕНС, ИНСТРУМЕНТ) като концептуални релации се счита за добър стил на дизайн на концептуален ресурс, защото позволява да се адресират систематично повечето от понятията, важни за областта, и свързва по естествен начин текстови и концептуал-

ни единици. В някои проблемни области като медицината и здравеопазването, естественият избор за извличане на понятия, описващи областта, е да се съпоставят *понятия* на важни *термини* (най-често съществителни), но относно дефинирането на релациите между тях може да има разнообразни подходи и съображения. Това ясно се вижда в най-голямата колекция от медицински термини UMLS.

Някои от експертите в тази област се обединяват около мнението, че концептуалните връзки между понятията могат да бъдат разпознати чрез (*i*) автоматично идентифициране на лингвистични релации между съответните термини в текстовото описание и чрез (*ii*) филтриране и специфициране на лингвистични релации в процеса на тяхната интерпретация като концептуални релации. Прилагайки тези принципи, ние предлагаме алгоритъм за автоматично извличане на релации между понятия чрез анализиране на дефиниции на термини. Служим си със съществуващи ресурси на английски език - източниците на UMLS.

2.3.2 Дефиниция на задачата

Задачата, която си поставяме, е да обогатим терминологичния речник, построен на предходния етап (Секция 2.2) с връзки между понятията, както е показано на Фигура 2.7. Така ще получим по-адекватен модел на проблемната област, който да подпомага извличането на информация от епикризи на български език. За целта използваме подход, при който обработваме дефиниции на термини от речника и разпознаваме автоматично връзки между понятия, споменати там. В последствие пренасяме тези връзки върху съответните термини в нашия речник. Анализът е фокусиран върху извличането на йерархични релации **IS-A** и релации от тип **AFFECTS**. Входният речник е на български, както и крайният резултат, а извличането на дефинициите и тяхната обработка се извършва на английски. Източник на дефинициите са ресурсите на UMLS.



Фигура 2.7: Процес на създаване на модела на предметната област.

2.3.3 Ресурси

Избор на релации

Използвахме семантичната мрежа на UMLS, за да подберем релациите между концепции, които са важни в биомедицината и са интересни за моделиране от гледна точка на OEE. В UMLS Semantic Network има мрежа от 52 релации, които изразяват експертна гледна точка за това как могат да бъдат свързани помежду си медицинските понятия. Въпреки че много малка част от тези връзки са експлицитно означени в UMLS, семантичната мрежа улеснява до известна степен задачата за извличане на релации, като ни помогна да прецизираме фокуса на това изследване. Подбрахме йерархичната релация **IS-A** и релациите от тип **AFFECTS**, които означават връзката между заболявания и органите, върху които те причиняват патологични изменения.

Дефиниции

Източникът на дефиниции на термини за това изследване е UMLS, а термините, чиито дефиниции обработваме, са тези от терминологичния речник на български език, създаден по метода отдолу-нагоре (описан в Секция 2.2). Тъй като нашият речник е на български език, а речниците, които ползваме от UMLS, са на английски, правим предварителна обработка на термините, за да можем

да извлечем съответните им дефиниции на английски език.

Терминологичен речник

Терминологичният речник съдържа понятията, които се срещат в описанията на крайници в колекция от 1000 пациентски записа на диабетно болни и е обогатен със синоними и парафрази. Той съдържа 165 термина. Превеждаме термините на английски език и където е възможно даваме по повече от един превод на термин, като се придържаме към лематизираната форма. Тези от тях, които означават усложнения на диабета, са преведени спрямо диабетната онтология OGMD[7] от BioPortal[8]. Стараем се да дадем повече от един лексикален запис на преведените термини (напр. “problem of the eye” and “eye problem”), за да се повиши вероятността за намиране на дефиниция на дадения термин при заявка към UTS сървъра. Списъкът с преведени термини се състои от единични думи и от фрази. Всеки от тези термини се изпраща като заявка към UTS Server, като се ползва алгоритъмът за търсене на точно съвпадение (описан в секция 2.2.4). Ако намери съвпадение за този термин в базата на UMLS, сървърът връща дефиницията на понятието и идентификатора му CUI (а понякога и повече от един идентификатор/понятие и съответно дефиниция, някои от които са извън фокуса на това изследване). От всички извлечени понятия ръчно филтрираме подходящите за изследването.

Множеството от дефиниции, извлечени по този начин, е основният ресурс за последващия анализ. Предимството на дефинициите от UMLS е, че те са в сбита форма - кратки параграфи текст, съдържат съществена информация за понятието и са написани на език, който е специфичен за предметната област (за разлика от напр. Wikipedia, където дефинициите са написани на популярен стил). Освен това дефинициите, които са от един и същи UMLS ресурс, са изказани по подобен начин.

[7]Ontology of Glucose Metabolism Disorder http://bioportal.bioontology.org/ontologies/OGMD
[8]BioPortal http://bioportal.bioontology.org/

RelEx

Друг инструмент, който ползваме, е RelEx - синтактичен и семантичен анализатор. Той обработва текст на английски език и връща като резултат депендентните синтактични структури, които описват изреченията и семантичното обвързване между обектите, базирано на синтактични и семантични категории. Главният компонент извлича депендентните връзки, допълнителни модули изпълняват функция за разпознаване на анафора и семантично обвързване. Ядрото на програмата е Link Grammar Parser[9] [50] [32] [19], който осъществява синтактичния анализ. От WordNet [37] [16] се вземат: основната морфология на английски, единствено число на думи, леми на прилагателните и инфинитиви на глаголите.

2.3.4 Метод

Извличане на дефиниции

Текстовият корпус в нашия експеримент е колекция от дефиниции на термини, извлечени от UMLS Metathesaurus. Всяко понятие с уникален CUI представлява едно значение на някой термин и има дефиниция в някой от речниците на UMLS (измежду стотиците ресурси, интегрирани в UMLS). Понякога едно понятие може да няма текстова дефиниция. Всеки термин в UMLS може да е етикет на едно или повече понятия (понякога до 5), а някои понятия може да имат до 8 дефиниции в различните речници; очевидно не всички те са от интерес за този анализ.

Ние ползваме сървърите за отдалечен достъп до UMLS – UMLS Terminology Services (UTS) и извличаме всички възможни дефиниции за термините, които наблюдаваме. Това е стандартна процедура за UTS. Таблица 2.5 представя някои дефиниции и понятия извлечени при изпращане на термина “pulse” като заявка към UTS. Резултатът съдържа идентификаторите на понятията, в този случай C0391850 and C0034107, съответните им етикети ”Physiologic pulse ” and ”Pulse taking ” и техните дефиниции.

Първото понятие има само една дефиниция, а второто има 2 дефиниции,

[9]Link Grammar Parser, http://www.link.cs.cmu.edu/link/

извлечени от различни UMLS източници. От извлечените дефиниции експерт филтрира само онези, които са от интерес за този анализ, т.е. отнасят се за анализираното подмножество от термини. Например от Таблица 2.5 са подбрани само 2 дефиниции: *Def 1* за "Physiologic pulse " и *Def 1* за "Pulse taking ", но *Def 2* е премахната, защото не се отнася до описанието на крайниците. Там, където дефинициите съдържат повече от едно изречение, ние анализираме само първото (предполагаме, че то носи най-важната информация).

Термин	**CUI**	**UMLS термин**	**Дефиниции**
pulse	C0391850	Physiologic pulse	*Def 1. The rhythmic wave within* the arteries occurring with each contraction of the left ventricle.
pulse	C0034107	Pulse taking	*Def 1.* The rhythmical expansion and contraction of an ARTERY produced by waves of pressure caused by the ejection of BLOOD from the left ventricle of the HEART as it contracts. *Def 2.* Actions performed to measure rhythmical beats of the heart.

Таблица 2.5: Извлечени дефиниции от UMLS за термина "pulse".

Трансформации на повърхностни характеристики

Прилагаме предварителна обработка на повърхностни характеристики на дефинициите, с което целим намаляване на възможностите за грешка в последващите фази. Характеристики като ортографията и пунктуацията са от съществено значение за автоматичния синтактичен анализ и малки трансформации в тях могат да подобрят резултатите от анализа. Филтрирането на информация, която не се отнася до извличането на релации между понятията, помага да се фокусираме върху съществената част от дефиницията, която носи информация за връзките между понятията. За тези трансформации прилагаме правила, които са разработени след анализиране на тренировъчния корпус от гледна точка на конкретната задача.

- първата буква на всяко изречение се трансформира в главна;
- на края на всяко изречение се поставя точка, ако тя липсва;
- примерите са премахнати от дефинициите;
- специални маркери, които означават източниците на информация са премахнати;
- транскрипциите за произнасяне на термините се премахват.

Така резултатите от синтактичния анализ на стъпка 1 се подобряват значително. Примери са показани на Таблица 2.6.

Оригинална дефиниция в UMLS	**Коригирана дефиниция**
(eh-DEE-ma) Swelling caused by excess fluid in body tissues.	Swelling caused by excess fluid in body tissues.
A blood vessel that carries blood away from the heart. **(NCI)**	A blood vessel that carries blood away from the heart.
swelling from excessive accumulation of serous fluid in tissue.	**S**welling from excessive accumulation of serous fluid in tissue.
The controlled release of a substance by a cell. **[GOC:mah]**	The controlled release of a substance by a cell.

Таблица 2.6: Нормализация на дефинициите, извлечени от UMLS

На избраните дефиниции правим синтактичен и семантичен анализ със системата RelEx, описана в Секция 2.3.3. Тя работи по следния начин:
Стъпка 1: Изпълнява Link Parser и превръща резултатите от него в структурирани характеристики, които се използват на следващите стъпки;
Стъпка 2: Изпълнява серии от алгоритми, работещи на ниво изречение, които модифицират характеристиките, извлечени на първа стъпка;
Стъпка 3: Минава по характеристиките, извлечени на предните стъпки, и ги преобразува до разбираем за потребителя финален резултат - дependентни структури и релации за семантично обвързване.

Дependентните парсери и link-парсерите са полезни за генериране на графи, които имат съответствие с концептуални графи и представят семантичните връзки между обектите. Ние използваме дependентните дървета и релациите от семантичното обвързване между обектите, които RelEx извежда, след об-

работка на дефинициите. Комбинираме ги в концептуална структура, чиито характеристики помагат за извличане на нови релации.

Трансформации на синтактичните структури

На стъпка 1 RelEx среща трудности при анализа на твърде прости или твърде сложни изречения от дефиниции. Само в 34% от дефинициите RelEx разпознава всички думи и обработва напълно входния низ. В останалите 66% поне една дума не е разпозната. Тъй като депендентният синтактичен анализ започва от глагола на изречението, за дефинициите, които са безглаголни, не може да се изведе синтактично дърво, както и да се направи семантично обвързване на обектите. Такива дефиниции са показани в Пример 2.3.1.

Пример 2.3.1. *(LIMB) A body region referring to an upper or lower extremity.*
(MUSCLE) One of the contractile organs of the body.
(PHALANX OF HAND) A bone of the hand.

Справяме се с този проблем, трансформирайки безглаголните фрази до изречения, като добавяме в началото на изречението термина, който дефиницията описва (в Пример 2.3.1 това са думите в скоби) и след него добавяме глагола "to be". Така изреченията от Пример 2.3.1 се преобразуват както е показано на Таблица 2.7.

Оригинална дефиниция в UMLS	**Коригирана дефиниция**
(LIMB) A body region referring to an upper or lower extremity.	Limb is a body region referring to an upper or lower extremity.
(MUSCLE) One of the contractile organs of the body.	Muscle is one of the contractile organs of the body.
(PHALANX OF HAND) A bone of the hand.	Phalanx of hand is a bone of the hand.

Таблица 2.7: Коригиране на безглаголни дефиниции, извлечени от UMLS

Окончателният резултат от RelEx включва конституентни и депендентни синтактични дървета (включително множеството от всички депенденти релации и характеристики), синтактично дърво от Link Grammar и релации, уста-

новени чрез правилата за семантично обвързване.

Фигура 2.8 съдържа резултатите от анализа на RelEx за конституентен парсинг и LingGrammar парсинг върху изречението:

> *Diabetic Cataract is a rare, usually bilateral, opacity shaped like a snowflake, affecting the anterior and posterior cortices of young diabetics.*
> *Диабетната катаракта е рядка, обичайно представлява двустранно помътняване, оформено като снежинки, засягащо предната и задната повърхност (на лещата) при млади диабетици.*

Под тях е изобразено депендентното дърво на изречението.

Добавяне на нови релации между понятията в модела

Релация IS-A

Извличаме нови релации от тип **IS-A**, като се базираме на депендентия граф на изреченията и следвайки общоизвестни правила за композиция на смисъла. Вземаме предвид обичайната структура на английското изречение *подлог-сказуемо-допълнение* (ПСД) и фактът, че дефинициите са кратки изречения. В Пример 2.3.1 са показани безглаголни дефиниции, които трансформираме чрез добавяне на глагола "to be" до варианта в Таблица 2.7. Изреченията с такава структура са много подходящи за автоматично извличане на релации **IS-A** от тях. Оказа се, че 10% от дефинициите в нашите учебни данни имат точно такава структура.

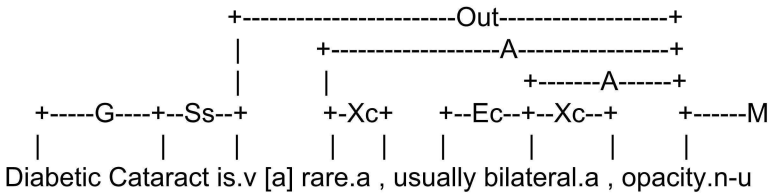
След анализ на специфичните характеристики на дефинициите от учебния корпус, създадохме следните правила, които позволяват автоматично да изведем релации от тип **IS-A**.

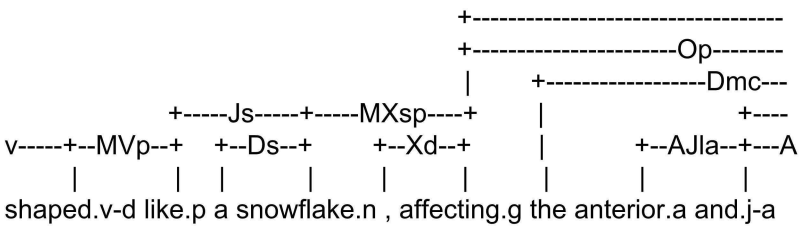
Правило 2.3.1. *Ако в едно изречение имаме депендентни релации, съответстващи на подлог (subj), сказуемо (глаголът "to be" в сегашно време), и допълнение (obj) и/или модификатор (nn) на подлога в главното изречение както следва:*

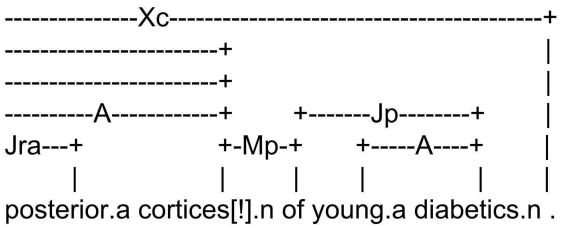
Constituency parse

(S (NP Diabetic Cataract) (VP is [a] (NP (NP (ADJP rare ,)
(ADJP (ADVP usually) bilateral ,) opacity) (VP shaped (PP like (NP a snowflake
(VP , affecting (NP (NP the (ADJP anterior and posterior) cortices)
(PP of (NP young diabetics))) .)))))))

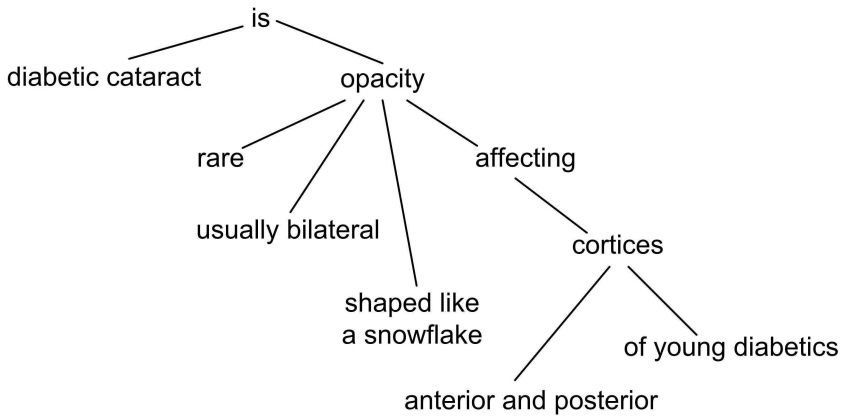
LinkGrammar parse







Dependency tree



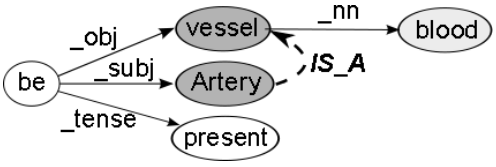
Фигура 2.8: RelEx резултати, илюстриращи депендентните връзки, които подпомагат извличането на релации.

obj(be, N3)
subj(be, N1)
tense(be, present)
nn(N1, N0)
nn(N3, N2),

тогава можем да заключим, че N1 **IS-A** *N3. Последните две NN-синтактични връзки не е задължително да са налични, но ако е налично N0, то ще е модификатор на подлога N1, а N2 ще е съответно модификатор на допълнението N3.*

В Пример 2.3.2 е показана дефиниция, което се обработва с Правило 2.3.1. Релациите, резултат от анализа на RelEx, и правилото за извличане на **IS-A** релации са показани и на графа от Фигура 2.9. С непрекъсната линия са изобразени релациите от RelEx, а новооткритата **IS-A** релация е маркирана с прекъсната.

Пример 2.3.2. *"Artery is a blood vessel that carries blood away from the heart." Артерията е кръвоносен съд който отвежда кръвта от сърцето.*



Фигура 2.9: Депендентен граф, обясняващ извличането на релация **IS-A** по правило 2.3.1

Релация AFFECTS

Другата важна релация е **AFFECTS** (засяга). Фокусираме се върху случаите, в които тя изразява връзка между усложнения/болести/симптоми и органа, който те засягат. Разглеждаме примерите, извлечени от дефиниции на усложненията на Diabetes Mellitus. За да извлечем релациите **AFFECTS**, представяме всяка дефиниция в лексико-семантична структура. Тази структура се построява на 5 стъпки, както е описано долу за изречението "*Diabetic Cataract is a rare,*

usually bilateral, opacity shaped like a snowflake, affecting the anterior and posterior cortices of young diabetics."(Диабетната катаракта е рядка, обичайно представлява двустранно помътняване, оформено като снежинки, засягащо предната и задната повърхност (на лещата) при млади диабетици.).

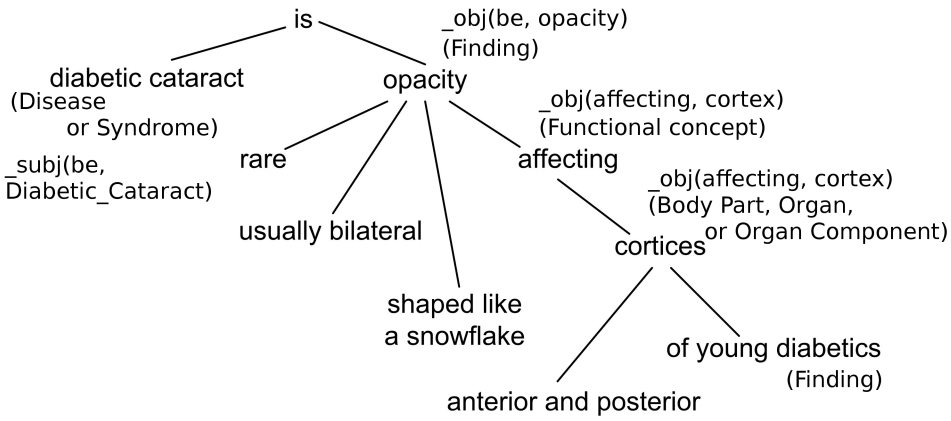
Анализът започва с обработка на дефинициите с RelEx - синтактичен и семантичен анализ.

Стъпка 1: В резултат на синтактичния и семантичен анализ RelEx извлича обектите, които се срещат в дефинициите. Създаваме списък от тези обекти.
Обекти: *{Diabetic Cataract, opacity, snowflake, cortex, affecting, Diabetic, diabetic, like}*

Стъпка 2: Разширяваме всеки обект с неговите възможни модификатори, извлечени при синтактичния анализ. Това могат да са прилагателни или съществителни в ролята на прилагателни, които формират група на съществителното, или словосъчетания, свързани с предлог, извлечен от парсера (напр. accumulation of amount). Пропускаме модификаторите, които са стоп-думи (напр. and, or, due etc.).
Обекти, разширени с модификатори, където е възможно: *{Diabetic_ Cataract, opacity, bilateral opacity, rare opacity, snowflake, shape like snowflake, cortex, and cortex, affecting, Diabetic, diabetic, young diabetic, like}*
От този списък премахваме следните обекти: “and cortex”, защото and е стоп-дума; “Diabetic”, защото вече е в списъка. Низът Diabetic_Cataract се преобразува до Diabetic Cataract.

Стъпка 3: Търсим в UMLS думи и словосъчетания от Стъпка 1 и Стъпка 2, за да проверим кои от тях са медицински термини. Извличаме термините, които имат най-близко разстояние по Hemming до началния обект.
Обекти, разпознати от UMLS като медицински термини: *{Diabetic Cataract (Diabetic Cataract), Retinal opacity (opacity), Snowflake retinal degeneration (snowflake), Visual Cortex (cortex), affecting (affecting), diabetic (diabetic)}*

Стъпка 4: Към всеки термин, извлечен на Стъпка 3, присвояваме неговия семантичен тип от UMLS.



Фигура 2.10: Интеграция на характеристики и извличане на нови релации.

Всички обекти от преходната стъпка с прикачен семантичен тип (в скобите): *{Diabetic Cataract (Disease or Syndrome), Retinal opacity (Finding), Visual Cortex (Body Part, Organ, or Organ Component), snowflake (Acquired Abnormality), affecting (Functional concept), diabetic (Finding)}*

Стъпка 5: Създаваме ново лексико-семантично представяне на дефиницията, при което всеки термин е представен чрез семантичния си тип.
Лексико-семантично представяне: *(Disease or Syndrome) is a rare, usually bilateral, (Finding) shaped like a (Acquired Abnormality), affecting the anterior and posterior (Body Part, Organ, or Organ Component) of young (Finding).*

Стъпка 6: Извличаме релация **AFFECTS** и нейните аргументи на база на характеристиките, които сме натрупали на предишните стъпки. Характеристиките, събрани на тези стъпки, ни дават възможност на стъпка 6 да работим с конкретни медицински термини, да имаме техните модификатори и депенденто представяне, което ни помага да различим понятието, съответстващо на засегнатия орган. На тази стъпка интегрираме всички събрани дотук характеристики, както е показано на фигура 2.10, и анализираме шаблони за извличане на релациите на тази основа.

Правило 2.3.2. *Ако имаме дependентна и семантична структура, които удовлетворяват следните условия:*
1. subj(be, N1) AND N1 hasClass Disease or Syndrome
2. obj(be, N2) AND N2 hasClass Disease or Syndrome or Finding
3. A verb V singling AFFECT relation
4. obj(V, N3) AND (N3 hasClass (Body Part, Organ, or Organ Component) OR the augmented entity of N3 identifies hasClass Body Part, Organ, or Organ Component ние можем да заключим по това правило, че N1 AFFECTS N3 и по Правило 2.3.1, че N1 **IS-A** *N2. Напр. Diabetic Cataract* **IS-A** *opacity AND Diabetic Cataract* **AFFECTS** *cortices .*

2.3.5 Дискусия

В този експеримент обработихме 194 дефиниции, които съответстват на 129 термина. Лексикални конструкции, които се отнасят до части на тялото и органи бяха намерени в 57% от дефинициите. Анализът на конституентните дървета показа, че тези термини (органи/части на тялото) са винаги в предложна фраза, приложена към глагола, който анализираме. Разгледахме шаблоните, които покриват релацията **AFFECTS** и направихме списък с глаголи, които изразяват наличието на тази релация. Лексикалните им форми могат да са в деятелен или страдателен залог. Най-често употребяваните лексикални форми са: affecting, consisting of и characterized by. Фразата "resulting from"сигнализира наличието на друга болест или условие, което предизвиква наблюдаваната болест.

Релацията **IS-A** е извлечена с точност 81%, докато в подмножеството от дефиниции, на които коригирахме синтактичната структура (Пример 2.3.2) резултатът беше 89%. Има няколко причини за сравнително ниската точност върху цялата колекция. Първата е некоректни синтактични дървета, в резултат от анализа, които се дължат на сложна синтактична структура. Друга причина е частичното разпознаване на словосъчетания, които съвпадат с подлога и/или допълнението и не позволяват правилото да се изпълни докрай. Покриваемостта на дефинициите, в които експлицитно е изразена **IS-A** релация е 86%.

Извличането на релацията **AFFECTS** е с по-малка успеваемост отколкото

релацията **IS-A**. Точността на извлечените релации е 24%. Един от факторите за този сравнително нисък резултат е често срещана структура на дефинициите, с която нашите шаблони съвпадат и е трудно да бъдат филтрирани. Това са случаите, в които обект с тип "*Disease or Syndrome*" е не конкретно наименование на заболяване, а типа на заболяването (например "acute infectious disease"). Релациите от този тип са вярни, но те не ни предоставят достатъчно информация за конкретно заболяване. Други причини са свързани с това, че изразяването на релация **AFFECTS** има много по-голяма вариативност; разстоянието между аргументите на релацията може да е сравнително голямо и това създава предпоставки за грешки; някои от думите, които сигнализират **AFFECTS**, са многозначни и реферират към други релации. Неправилната пунктуация и сложни именни фрази като показаните в Таблица 2.8, също са причина за грешки в синтактичния анализ. Понякога парсерът не успява да разреши многозначността и обърква глагол със съществително за думи като *measure*, *joint*, и др. Таблица 2.8 показва и грешен синтактичен анализ, където съществителното *touch* е било анализирано като глагол. Имайки предвид ограничения брой на терминологията, която разглеждаме, считаме, че този подход е подходящ за подпомагане на експертите в процеса на обогатяване на модела на предметната област, където покритието на предложените релации е от по-голямо значение.

Изречение	**Конституентно синтактично дърво**
Touch; the faculty of touch, the sensation produced by pressure receptors in the skin.	(S (***VP touch*** [;] (NP (NP the faculty) (PP (NP of touch) , (NP (NP the sensation) (VP produced (PP by (NP pressure receptors))(PP in (NP the skin))) .)))))
A joint connecting the lower part of the femur with the upper part of the tibia.	(S [a] (S (***VP joint*** [connecting] (NP (NP the lower part) (PP of (NP the femur))) (PP with (NP (NP the upper part) (PP of (NP the tibia)))))) .)

Таблица 2.8: Грешно разрешаване на многозначността на ниво част на речта

2.3.6 Резюме

Нашата цел с това изследване е да проверим наличността на ресурси за OEE и тяхната готовност да послужат за разработването на нови концептуални ресурси. Представихме един подход, който прави възможно автоматичното извличане на релации между медицински понятия чрез повторно използване на съществуващи ресурси и програми за OEE и чрез прилагане на допълнителни правила за трансформация. Като една по-далечна задача си представяме например извличане на енциклопедична и справочна информация от UMLS за учебни цели. Може би полу-автоматичното построяване на пълен концептуален модел с етикети на български език, който изброява засегнати органи за дадени заболявания, би било интересна задача.

При анализиране на резултатите от междинните стъпки забелязахме, че често причината нашият алгоритъм да не успее са в грешни синтактични дървета, изведени от RelEx. По-добър парсинг би помогнал за по-успешно разрешаване на задачата. Някои от **IS-A** релациите, които извлякохме, бяха вече налични в UMLS Metathesaurus, но около 45% бяха новосъздадени. Само 3 от извлечените релации **AFFECTS** бяха налични в Metathesaurus и те бяха конкретни, но от типа "*has relationship other than synonymous*", "*narrower*", or "*broader*" . В бъдеще възнамеряваме да направим по-задълбочено оценяване на правилата за извличане на релации и да работим над правила за специфични подтипове като "*functionally related to*", "*manages*", "*treats*", "*disrupts*", "*complicates*", "*interacts with*", "*prevents*".

Глава 3

Структуриране на текстови описания в биомедицината

Пациентските записи са богат източник на информация относно здравното състояние на пациента и неговото лечение през годините на заболяването, но обикновено те съществуват само във формата на свободен текст. Последните години се влагат сериозни усилия във водещи медицински институции по света за структурирането на тези данни и предоставянето им за по-нататъшна автоматична обработка, така наречената *повторна употреба на ПЗ*. Използването на записите ще допринесе за напредък в персонализирането на лечението и за откриването на явления, които могат да се наблюдават над голям обем от данни, и като цяло за подобряването на здравеопазването и здравния мениджмънт.

Автоматичното структуриране на текстови описания като ПЗ става възможно с прилагане на техниките за ОЕЕ. При ОЕЕ се извлича висококачествена информация от текст с изработване на шаблони и наблюдаване на тенденции със средствата на статистическите методи за учене на шаблони. Наблюдават се разнообразни лингвистични, структурни и повърхностни характеристики на текста, извличат се късове от информация спрямо предварително определени шаблони, след което тази информация се интерпретира и нормализира, свързва се със съществуващи понятия в контекста на предметната област.

ПЗ на български език имат комбинация от характеристики, които ги правят привлекателни за разнообразни научни задачи от гледна точка и на медицината, и на езика. Те са ценен източник на данни за здравеопазването, съдържат

богата биомедицинска терминология и в същото време са на език, за който съществуват сравнително малко ресурси за ОЕЕ. В тази глава представяме подходи за структуриране на характеристики, симптоми и описания на състоянието на пациента от медицински епикризи на български език, с помощта на техники за ОЕЕ.

3.1 Извличане на симптоми на диабет от пациентски записи

3.1.1 Въведение

Следвайки тенденцията за повторно използване на пациентски записи, ние разработихме методи за извличане на симптоми на диабета от медицински епикризи на български език. Целта ни е да структурираме свободния текст чрез автоматична обработка на естествен език. Понякога в болничните практики в България, лабораторните данни в ПЗ са налични само в текста, където лесно могат да бъдат разчетени от човек, но за да бъдат анализирани автоматично е необходимо да бъдат направени допълнителни обработки. Тук представяме алгоритъм, който съчетава техники за машинно самообучение и анализ, базиран на правила, с който разпознаваме автоматично симптоми на диабета - фрази и парафрази, за които няма "канонични форми", описани в речник. Анализираме ПЗ на пациенти, диагностицирани с диабет. Извлечената информация е ценна, защото е от една страна много важна за лекарите и от друга страна, трудно може да се наблюдава директно в голяма колекция от неструктурирани записи, поради богатството и вариативността на изказните средства в естествения език.

3.1.2 Дефиниция на задачата

Фокусираме се върху извличане на *нива на кръвна захар*, *промяна в телесното тегло* и *полиурично-полидипсичен синдром*, които са едни от доминиращите фактори при поставяне на диагноза за диабет, както и възрастта на пациента. При изследването се водим изцяло от наличните корпуси от данни, от които извличаме речници със сигнализиращи изрази и правила за извличане на състоянията. Целта ни е също така да сравним успеваемостта на подходите базирани на машинно самообучение и тези на правила, при класифициране на изреченията, спрямо това дали съдържат симптомите, които ни интересуват. На Пример 3.1.1 са показани изречения, в които се срещат тези симптоми, изказани по различни начини.

Пример 3.1.1.

** При изследване **кръвната захар е била - 14 mmol/l.***

** От 1г е сменена пероралната терапия с Амарил, понастоящем 6мг и Авандия, понастоящем 8мг, като на този фон **достига стойности на КЗ до 15 ммол/л**.*

** През октомври месец т.г. във връзка с несъответно **повишен на кръвната захар инсулин** е включено лечение с метформин /Сиофор/ с постепенно увеличение на дозата от 2х1/2т. до 3х1т.*

** Повод за настоящата хоспитализация е изтръпване и мравучкане на крайниците с намалена чувствителност и **лош кръвно-захарен контрол**. * Постъпва по повод на **полиурично-полидипсичен синдром**, **редукция на теглото** и кетоацидоза.*

** От дълги години **с наднормено тегло** във времето **развила затлъстяване**.*

** От откриването на диабета до настоящия момент с диетичен режим е **редуцирала килограмите си с около 40**.*

** Със **затлъстяване I степен от дълги години**, установена чернодробна стеатоза и дислипидемия, провеждала е лечение със статино и фибрати.*

Считаме, че същият подход може да се приложи при разпознаване на други симптоми или изрази, свързани със заболяването, които имат сходна структура на описанието.

3.1.3 Ресурси

Пациентски записи

Това изследване е направено върху свободния текст от ПЗ на пациенти, диагностицирани с диабет в Болницата по ендокринология към Медицински университет София. Описанията, които се стремим да разпознаваме автоматично са налични в секцията *История на заболяването* (анамнеза), където е записано развитието на диабета, усложненията, тяхното съответно лечение и др. Описанието на един симптом се намира в рамките на едно изречение, а понякога и повече от един симптом може да бъде описан в същото изречение, както е показано в Пример 3.1.1.

Езикът на епикризите има своя специфика, която трябва да се вземе предвид при автоматична обработка на текста. Изследвахме дали документите от-

говарят на условията за специализиран език. Установихме, че те притежават особености, подобни на тези, които се забелязват в специализираните езици, произхождащи от английски [55]. Такива характеристики са лексикалното затваряне (виж в речника), затваряне по отношение на част на речта за думите в текста, затваряне по отношение на типа на изреченията като последователност от морфологични маркери. За разлика от предишни изследвания на тази тема като [35] и [54], затваряне по типове изречения беше демонстрирано за първи път. Тези характеристики се установяват с функции на насищане. При добавяне на нови документи броят на новите маркери за лексикални (морфологични, типове изречения) се увеличава сравнително пропорционално до даден момент, след което функцията се насища и с прибавянето на нови документи не се забелязват нови характеристики от съответния тип.

Освен това, епикризите съдържат медицинска терминология, изписана с латински символи (около 1% от всички думи в текущия корпус), понякога се срещат латински термини, изписани с кирилица. Налични са специфични съкращения и абревиатури на български и на латински (около 3% от думите), числа (около 16% от низовете) и др. В епикризите завършените глаголни изречения са рядкост, тъй като текстът съдържа основно изречения-фрази. Понякога няма съгласуване между частите на изречението и пунктуацията не е правилна/налична. Българската морфология е силно флективна, термините се срещат в текста в голямо разнообразие от лексикални форми, което допълнително усложнява автоматичната обработка. Срещат се дори членувани имена на лекарства, например "*инсулинът*".

Нашият учебен корпус е подмножество от изречения от анамнези, които са свързани с описание на симптоми. Те са анотирани на ниво изречение с маркери за типа на симптома, който съдържат, като на ниво дума са направени BIO (begin-inside-outside) анотации, маркиращи границите на описанието на симптома. Примери са посочени на Таблица 3.1 и 3.2

Изреченията са извлечени от 100 епикризи. Изреченията, съдържащи описание на кръвна захар, са анотирани с клас BLOOD_SUGAR (съответно B-BLOOD_SUGAR, I-BLOOD_SUGAR), тези описващи промяна в теглото, с WEIGHT_CHANGE (B-WEIGHT_CHANGE, I-WEIGHT_CHANGE), описващите полиурично-полидипсичен синдром с клас PP_SYNDROME (съответно

полидипсо-полиуричен синдром	
Заболяването	<OTHER>
е	<OTHER>
диагностицирано	<OTHER>
преди	<OTHER>
6	<OTHER>
години	<OTHER>
във	<OTHER>
връзка	<OTHER>
с	<OTHER>
полидипсия	<B-PP_SYNDROME>
.	<OTHER>

Таблица 3.1: BIO анотация за полидипсо-полиуричен синдром.

B-PP_SYNDROME и I-PP_SYNDROME), възраст с клас AGE (съответно B-AGE и I-AGE) и всички останали изречния са маркирани като "други" с клас OTHER. Изреченията, които съдържат повече от един симптом, са анотирани с повече от един етикет. Тези данни са ползвани, за да се научат правилата и речниците, които се използват за разпознаване. Експерименталното множество от данни съдържа 2031 анотирани изречения. Документите са сегментирани на изречения и думи. За да се преодолее инфлексията и да се постигне по-широко покритие на правилата ние ползваме корените на думите, отделени чрез програма стемер за български език [38].

Речници

Алгоритъмът разчита на няколко списъка, които са построени ръчно от анотираните тренировъчни данни. Те са следните:

(***i***) речник $P1$ със симптоми;

(***ii***) речник $P2$ с ключови думи;

(***iii***) речник $P3$ с "гранични изрази" отляво;

(***iv***) речник $P4$ с "гранични изрази" отдясно;

(***v***) речник $P5$ с изрази, сигнализиращи отрицание;

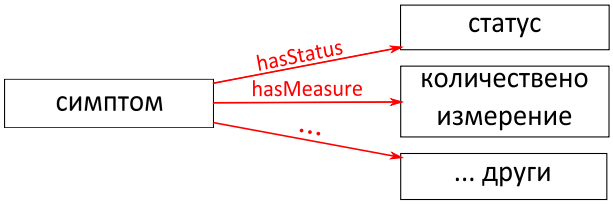
промяна в теглото	
През	<OTHER>
последната	<OTHER>
1.5	<OTHER>
г.	<OTHER>
е	<OTHER>
наддала	<B-WEIGHT_CHANGE>
с	<I-WEIGHT_CHANGE>
около	<I-WEIGHT_CHANGE>
15	<I-WEIGHT_CHANGE>
кг	<I-WEIGHT_CHANGE>
на	<OTHER>
фона	<OTHER>
на	<OTHER>
лечение	<OTHER>
с	<OTHER>
Манинил	<OTHER>
.	<OTHER>

Таблица 3.2: BIO анотация за промяна в теглото.

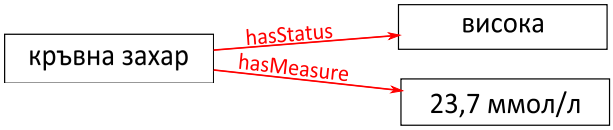
(*vi*) речник $P6$ с изрази, означаващи статус на симптома.

Речникът със симптоми $P1$ съдържа думи и фрази, срещащи се в описанията на симптоми, напр. “гликемичен контрол”, “хипогликемия” и др. Той се използва за определяне на симптома. Всички отделни думи, които се срещат в този речник (напр. “гликемичен”, “контрол”, “хипогликемия”), без стоп думите, формират *речник с ключови думи* $P2$, който се използва във *фаза 1* - класификационна задача.

Други два речника $P1$ и $P2$ съдържат “гранични термини”: единият съдържа *граничните изрази отдясно на описанията*, а другият - *граничните изрази отляво на описанията*. Това са думи и фрази, отделящи описанията на *кръвна захар* или *промяна в теглото* от други наблюдения. И двата речника са сортирани по намаляваща стойност на вероятността им да се срещат като граничен



Фигура 3.1: Общ шаблон на отношенията, представляващи описание на симптом.



Фигура 3.2: Шаблон на релацията *ниво на кръвната захар*.

термин.

Създават се също и $P5$ - *речник с изрази, сигнализиращи отрицание* (напр. “няма данни за”, “не се отчита” и др.) и $P6$ - *речник с изрази, означаващи статус на симптома* (e.g. “добро”, “лошо”, “увеличено” и др.).

3.1.4 Методи

Стремим се да идентифицираме автоматично описанията на симптоми в анамнезата на ПЗ. Формално погледнато всяко описание на симптом е релация между наименованието на симптома, неговия статус (посока на отклонение) и количественото му измерение. Визуално тя може да се представи както на Фигура 3.1. На Пример 3.1.2 описанието се изразява в релацията между наименованието - *кръвна захар* и стойността на измерването - *14 mmol/l*. Визуално шаблонът на релацията може да се представи както на Фигура 3.2.

Пример 3.1.2. *При изследване* ***кръвната захар е била - 14 mmol/l.***

Намирането на симптоми в текста може да се разглежда като последователност от няколко подзадачи.

Фаза 1. Идентифициране на изреченията, съдържащи описания на симптомите, които ни интересуват.

Фаза 2. Идентифициране на симптома.

Фаза 3. Идентифициране на статуса на симптома.

Фаза 4. Идентифициране на стойностите, които се отнасят до разглежданото състояние и съответните мерни единици - mmol/l, kg и др.;

Фаза 5. Обработване на отрицанието.

Фаза 6. Идентифициране окончателните граници на описанието - налагане на граничните термини отляво и отдясно.

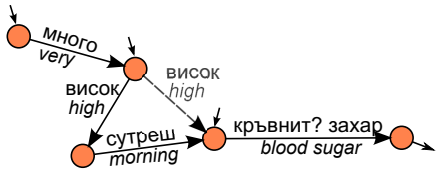
При анализ на резултатите ще реферираме към Фаза 1 и 2 като "фокус", в смисъла на фокусиране върху състоянието, от което се интересуваме.

Фаза 1. Идентифициране на изреченията, съдържащи описания на симптомите, които ни интересуват

За да се установят изреченията, съдържащи описание на симптоми на *Фаза 1*, бяха направени експерименти с два отделни подхода: подход с *машинно самообучение* и подход, *базиран на правила*.

При подхода с машинно самообучение обучаваме модел с бинарен класификационен алгоритъм. Експериментираме с 3 множества от характеристики на текста *(i)* всички думи от корпуса с изключение на пунктуация и числа; *(ii)* изразите от *речника с ключови думи* $P2$ и *(iii)* ограничено подмножество на *речника с ключови думи*. Във всички тези случаи изреченията се представят като документи и характеристичният вектор на всеки документ съдържа булеви стойности за всеки елемент. *Стойността на един елемент от вектора е* $true$*, само ако документът съдържа съответния токън/характеристика, иначе е* $false$. Например изречението "*Наблюдават се завишени нива на кръвната захар.*" ще има характеристичен вектор с $false$ по всички позиции, с изключение на тези за "завишен", "кръвна" и "захар", които са думи от речника $P2$. В тези позиции елементите на вектора ще бъдат $true$. Ползваме и експериментален корпус, който разделяме на части за обучение и тестване. Речниците, които ползваме, са научени от учебните данни. На тази фаза класифицираме изреченията в 2 класа - (i) такива, които съдържат симптоми, които ни интересуват и (ii) други.

В експримента, базиран на правила, съставяме регулярни изрази, които да съвпадат с описанията на симптоми в учебния корпус. Моделираме ги в прозо-



Фигура 3.3: Добавяне на нови връзки между корените.

рец от 5-7 думи наляво и надясно от централните термини, описващи симптома, в зависимост от типа на израза. Когато съставяме правила като на фигура 3.3 се въвеждат нови връзки между думите, които не са последователни в учебните данни, но биха имали смисъл употребени последователно. Връзките, отбелязани с непрекъснати линии са налични в учебния корпус, а прекъснатите линии са добавени ръчно. Този подход помага на процедурата за идентифициране, защото не ограничава правилата, извлечени от текста, а само ги разширява и то по контролиран начин. Освен това, тези правила са създадени от текстове с думи, сведени до корени, като по този начин частично се преодолява проблемът със съгласуването. Правилата имат по-широко покритие на възможните сигнализиращи думи. Някои от тях са показани на Таблица 3.3. Изреченията, които имат съвпадение с тези правила, се считат за съдържащи симптомите, които ни интересуват, и преминават за обработка на *Фаза 2.*

кръвн[аи](т)? захар
(незадоволителен|добър|лош|отлич) гликемич контрол
полидипсо-полиурич((синдром|оплаква))?
(\d)[1,3] (г|год)\.

Таблица 3.3: Фаза 1 - правила за последователности от думи след определяне на корените.

Фаза 2. Идентифициране на типа на търсения израз

Фаза 2 е изцяло базирана на правила, които се прилагат върху изреченията, подбрани на *Фаза 1.* Правилата са изградени над изрази, сигнализиращи наличието на описание на кръвната захар, промяна в теглото, полиурично-полидипсичен синдром или възраст. Определят се централните термини, опис-

ващи типа на симптома. В случая, когато разглеждаме система, изцяло базирана на правила, *Фаза 1* и *Фаза 2* съвпадат.

Фаза 3. Идентифициране на статуса на симптома

На *Фаза 3* се разпознава статусът на симптома. Нивото на кръвната захар най-често се цитира като *ниско*, *високо* или *нормално*, а може да бъде още *лошо* или *добро*, телесното тегло може да бъде *увеличено* или *намалено*. Думите в контекста, които сигнализират статуса на даден симптом, се намират отляво на централните термини в описанието на симптома, напр: *с високи стойности* на кр. захар; *лош* гликемичен контрол.

Фаза 4. Идентифициране на стойностите, които се отнасят до разглежданото състояние

Анализът на *Фаза 4* е свързан с динамично разширяване на десния контекст на описанията на симптоми от предишната фаза, с цел да се открият всички необходими атрибути на търсените релации. На тази фаза се стремим да идентифицираме количествено измерение на показателя, който ни интересува - стойността на кръвната захар в mmol/l или промяната в теглото в кг. Тези стойности се срещат в текста по няколко начина – като интервални стойности (напр. *между 4 и 6,5 mmol/l*); като максимална стойност, достигната в даден период (напр. *до 10-11 mmol/l*) или като конкретна стойност (напр. *5 mmol/l*). Разпознаването на стойностите става на два етапа. Първо прилагаме правила, които откриват думи от речника, сигнализиращи типа на представяне на стойността, напр. *между*, *до*, *над*, *около*. След това разширяваме контекста надясно от тази дума, и прилагаме правила за откриване на количествени стойности според типа, който сме определили. Така например, ако алгоритъмът разпознае дума, сигнализираща интервална стойност като *между*, той разширява контекста надясно до следващите 2 числа и мерна единица след втората, но разширяването става с не повече от 7 позиции. Ако числата останат извън този прозорец, те се игнорират и алгоритъма за определяне на стойност не успява. Прозорецът от 7 думи беше експериментално установен след анализ на учебните данни. В този прозорец обикновено се срещат глаголи, изразяващи темпорал-

ност, които се разполагат между централните термини и цифровото изражение на резултатите от лабораторни тестове (вж показано на Таблица 3.4).

кръвнозахарна стойност до **10-11 mmol/l**
стойностите на **кръвната захар** *са били* между **4 и 6,5 mmol/l**

Таблица 3.4: Разпознаване на резултати от лабораторни тестове.

Фаза 5. Обработване на отрицанието

Във *Фаза 5* обработваме отрицанието. В текста се наблюдава ограничено срещане на отрицания. Това се дължи на факта, че в българската клинична практика се описват предимно състоянията, в които се наблюдават патологични промени. Изразите, които сигнализират отрицание, се срещат в левия контекст на фразите, маркирани на *Фаза 1* и те модифицират изразите, разпознати на *Фази 2 и 3*. Например: *не съобщава за...*; *не [много] високи стойности на...*

Фаза 6. Идентифициране окончателните граници на описанието - налагане на граничните термини отляво и отдясно

На *Фаза 6* се уточняват окончателните граници на извлеченото описание. Те се определят от думите в контекста, които сигнализират началото на израза, края му и идентифицираните атрибути. Изразите, които разглеждаме започват или в началото на изречението, или следват друго описание на симптом и са свързани с него със съюз. Краят на израза или съвпада с края на изречението, или се сигнализира от количествена стойност или съюз, след който има описание на друг симптом (вж Таблица 3.5). Правилата за определяне на границите са приложени отдясно и отляво на вече идентифицираните атрибути, започвайки от правилата с най-висока вероятност и продължавайки в намаляващ ред, докато се намери съвпадение. Ако няма съвпадение в прозорец от 7 думи, за дясна граница се приема най-дясната дума от текущия израз и лявата граница на израза е или първата му дума, или отрицанието или статуса на състоянието.

Лява граница на описанията на симптоми
при кръвна захар...
на фона на лош гликемичен контрол...
с високи стойности на кръвната захар...
Дясна граница на описанията на симптоми
...кръвна захар - 14 mmol/l.
...лош гликемичен контрол и кетоацедоза.

Таблица 3.5: Лява и дясна граница на описанията.

3.1.5 Оценяване

Фаза 1 - Правила и машинно самообучение

Оценяваме нашия подход от няколко различни аспекта. Първо сравняваме класификацията на изречения, използвайки машинно самообучение с подхода, базиран на правила във *фаза 1*. Извършваме това сравнение само за симптомите "*ниво на кръвна захар*" и "*промяна в теглото*".

В случая с МС, всяко изречение се разглежда като отделен документ и е представено чрез булев характеристичен вектор. Стойността на атрибута x_i е *false*, ако изречението не съдържа съответната дума и е *true*, ако го съдържа. Напр. един от атрибутите в това множество е "захар", ако документът (изречението) съдържа този низ, стойността за атрибута "захар" е *true*, иначе е *false*.

В класификационната задача експериментирахме с няколко алгоритъма, между които Naïve Bayes, SVM, J48 и J48 даде най-добри резултати. J48 decision tree се счита за подходящ за класификация на бинарни данни [62]. За класификацията използваме Weka Data Mining Software [22]. Най-добрите резултати са показани на Таблица 3.6.

За да постигнем тези резултати направихме серия експерименти върху тестовото множество, използвайки характеристики, подбрани от учебните данни. В първия експеримент използвахме като множество от атрибути всички токъни от текстовата колекция освен стоп думите. Постигнатото хармонично средно $F-$*мярка* е около 82% при 10-fold cross-validation и за описания на кръвната захар, и за промяна в теглото.

Метод	Точност	Покриваемост	F-мярка
J48 кръвна захар 22 атр.	93.80	85.70	89.60
J48 промяна на теглото 16 атр.	94.30	85.30	89.60
Правила кръвна захар	96.40	90.00	93.09
Правила промяна в теглото	98.50	92.00	95.14

Таблица 3.6: 3. Оценяване на фаза 1. Резултати от подходите базирани на правила и на МС.

Данните са силно небалансирани, тъй като статусът на тези симптоми се изказва в 1 или най-много 2 от изреченията в анамнезата - само около 20% от изреченията съдържат описание на кръвната захар и само 6% промяна в теглото. Направихме балансирано учебно множество от равни по брой записи от всеки клас (отново обучавайки модели "един към всички') и постигнахме $F-$*мярка* до 89% за кръвната захар и за промяна в теглото.

В следващия експеримент множеството от атрибути съдържаше само елементи от *речника с ключови думи*. Това покачи резулатите при класификационната задача до 89.6% $F-$*мярка* за разпознаване на изречения, съдържащи описание на кръвна захар и промяна в теглото. Това са най-добрите постигнати резултати, които са показани в Таблица 3.6.

Таблица 3.6 показва, че точността на подходите, които са базирани на правила, е по-висока от тази, получена чрез класификация с машинно самообучение. Но при анализа на грешките забелязахме, че при този подход някои положителни примери са били случайно класифицирани като такива, защото въпреки че съдържат термини, сигнализиращи наличието на търсеното описание, съвпадението на правилата е станало с описания на други симптоми. По отношение на покриваемостта има разлика в резултатите, но и в двата случая можем да считаме, че те са приложими и за по-големи обеми данни, без да се използва знание за областта, особено в задачи като тази, където точността на извличане е по-важна от покриваемостта.

	Фаза	Точност	Покриваемост	F-мярка
кр. захар	Ф.1&2 Фокус	96,4	90,0	93,09
	Ф.3 Статус	91,00	45,50	60,60
	Ф.4 Стойности	88,90	77,80	83,00
	Ф.5 Отрицание	96,30	94,20	95,20
	Ф.6 Граници	97,00	96,00	96,50
тегло	Ф.1&2 Фокус	96,60	90,60	93,50
	Ф.3 Статус	86,20	78,10	82,00
	Ф.4 Стойности	87,50	70,00	77,80
	Ф.5 Отрицание	NA	NA	NA
	Ф.6 Граници	82,70	75,00	78,70

Таблица 3.7: 4. Резултати от отделните фази на прилагане на правилата

Оценяване на отделните фази

Резултати от отделните фази на анализа, базиран на правила, са показани на таблица 3.7.

На *Фаза 2* думите от учебните данни са напълно разпознати; има група от низове, които не са налични в учебните данни, но по време на обработката във *Фаза 1* съвпадат с правила, които разпознават дадените симптоми. Тези низове са добавени към описанието без да се конкретизира типът на техния статус. Някои думи не могат да бъдат идентифицирани по 2 причини – те не са налични в учебните данни или са разместени спрямо слотовете в шаблоните от учебните данни (напр. прилагателно е използвано след съществително, вместо да е преди него, както е и във всички учебни примери). 45% от атрибутите, които изразяват статус на кръвната захар, са разпознати и 78.1% от тези изразяващи промяна на теглото. Въпреки че покриваемостта на описанията за кръвна захар изглежда да е ниска на тази фаза, резултатите са всъщност добри за нас, защото по време на анализа на грешки ние открихме, че 60% от думите, които не са идентифицирани, всъщност са едни и същи.

На *Фаза 3* главният проблем за разпознаване на стойности са изразите, които съдържат и букви, и цифри, и представят резултатите от лабораторните тестове. Те се срещат сравнително рядко, имат голямо разнообразие в лекси-

калното представяне и грешки в записа и това е и причината за невисоката точност при разпознаването им. За да се научим коректно да разпознаваме такива примери, са необходими голямо количество учебни данни. Един подход за преодоляването на този проблем е предварително да се генерира списък с такива изрази, включващ букви, цифри и техните варианти. Отрицанието във *Фаза 4* е разпознато с висока точност.

На *Фаза 5* всички проблеми по определяне на границите на изразите за кръвна захар са разрешени успешно, с изключение на един. В учебните данни всички интервали, описващи стойности на кръвната захар, са от типа *от 12 до 14 mmol/l*, а в този случай той е бил записан като *от 12 mmol/l до 14 mmol/l*. Това довежда до грешно разпознаване на дясната граница и само частично разпознаване на стойностите на кръвната захар. Ако разполагаме с по-голям учебен корпус, този проблем би бил автоматично преодолян, а за неговото решаване би помогнала и една предварителна фаза на обогатяване на правилата с вариации на представяне, заимствани от други източници или генерирани от човек. Би било от полза за разпознаването на симптомите да добавим нови лексикални алтернативи и парафрази в допълнение на корените и фразите в регулярните изрази. При нашия подход изцяло се водим от учебните данни, защото целта ни е да видим колко далеч можем да стигнем в анализа на текста, без да ползваме друго знание за езика или предметната област.

Разширението на правилата, както е показано на Фигура 3.3, помогна при идентифицирането на описания на кръвната захар в два от случаите. Вярваме, че такива разширения при по-голям обем данни биха имали по-голямо положително влияние при извличането.

Извличането на описания за *полидипсо-полиуричен синдром* и *възраст на пациента* са направени с помощта на правила - регулярни изрази, базирани на лексикални, повърхностни и контекстни характеристики на думите, както и на речници със сигнализиращи думи, но не следват описаната горе последователност. Възрастта е извлечена с *F-мярка* 89,44%, а *полидипсо-полиуричен синдром* с *F-мярка* 90%.

3.1.6 Резюме

Предлагаме унифициран подход за разпознаване на медицински описания на симптоми, които имат смисъла на релации между няколко обекта - симптом, неговия статус и числово измерение, и са представени от широк набор от парафрази в свободните текстове на епикризите на български език. Резултатите показват сравнително висока точност при идентифициране на описания на симптоми в текстове на ПЗ. Това е постигнато с употребата на правила, извлечени от учебни данни и минимални допълнителни усилия за разширяване на покритието на правилата - свеждаме входните документи до корени на думи и ръчно добавяне на вариативност в правилата, по преценка на експерта. Разпознаване на класа на изреченията е направено с почти същата точност и при двата подхода - бинарен J48 класификатор и с правила при анализа на *Фаза 1* дори без да се включват ключови термини в класификацията. Тези резултати подсказват, че е подходящо ползването на автоматична класификация с МС за подобни задачи в бъдеще, поради гъвкавостта на подхода и сравнително малките усилия за анотиране на корпуса на ниво изречние.

Като продължение на това изследване ще опитаме да обобщим алгоритъма до по-абстрактно ниво, така че да бъде лесно приложим за идентифицирането на описания на други симптоми, лечение и др. Също така ще наблегнем на автоматичното извличане на правила чрез анализ на характеристичните вектори на изреченията и резултата от класификацията.

3.2 Разпознаване на събития и тяхната темпоралност в медицински текстове

3.2.1 Въведение

В областта на OEE се отделя сериозно внимание на изследване на времевия аспект. Когато извличаме връзки между обектите в текста, те често са валидни само за определен период от време и от правилното разпознаване на този период зависи правилното интерпретиране на данните. В медицината и обработката на медицински данни/текстове времевият аспект също има изключително голямо значение. Можем да демонстрираме предизвикателството, което представлява откриването на темпорална зависимост чрез следния пример: да се направи автоматично разграничение между споменавания на минали събития (напр. *Пациентът има история на...*) от настоящи. Това предизвикателство може да бъде много по-голямо и да обхваща не само семантиката на времевите изрази, но да включва цялостната структура на клиничните документи: например датата , на издаване на документа, продължителност на дадени събития, които предизвикват други и т.н. В тези случаи е необходимо да се анализират темпоралните релации между събития и времеви изрази в този конкретен текстов жанр.

В компютърната лингвистика се влагат значителни усилия за разрешаване на проблема как да се открива и третира времевият аспект. Един от най-важните въпроси в тази връзка е как да представяме темпорални релации - с теории, вариращи от предикатна логика от първи ред до теорията за референтната точка на Reichenbach [43]. Едни от най-важните форуми за представяне на съвременни изследвания в областта на темпоралността са състезанията TempEval. Най-близки до нашата тема са *Задача B от 2007*[1] и *Задача D през 2010*[2], където основните изискванията бяха да се поставят маркери на темпоралните релации между предварително дефинирани събития в документа и датата на създаване на документа. През 2007 бяха демонстрирани методи с машинно самообучение, системи, базирани на правила и хибридни системи, но нито една

[1]TempEval 2007 - http://www.timeml.org/tempeval/

[2]TempEval 2010 - http://www.timeml.org/tempeval2/

от системите не получи резултати, които съществено да се различават от другите. Средната точност на резултатите беше 0.76 [58]. Тази задача се повтори в Tempeval 2, този път една от системите постигна неочаквано ниски резултати (няма детайли за методологията), но повечето от другите системи постигнаха сходни резултати със средна точност 0.78 [59]. В [6] се описва описва система за определяне на темпорално кохерентни сегменти от клиничен текст и релации между тях, която използва машинно самообучение. Разглежда се много ограничено множество от три релации. Главното предимство на техния подход е, че той извежда обща подредба (global ordering), докато подходите в Tempeval извеждат по-скоро локализирани релации. [6] постига $F-$*мярка* 83% на задачата за сегментиране на времеви маркери и точност 78.3% на класификацията на трите типа релации. Този подход е доста амбициозен към разглеждането на сегментацията по темпорални маркери и постига високи резултати.

"Събитията" в медицинската терминология са широка категория с разнообразни дефиниции. В предишни i2b2 състезания[3], подобни задачи включваха извличане на медицински понятия, разпознаване на лекарства, разпознаване на състоянието на пациента и др., макар и с по-просто представяне[4]. Според [56], в състезанието i2b2 през 2010, системата, която е с най-добри резултати в извличането на понятия, използва условни случайни полета (CRF). В някои от подходите авторите разделят проблема на подзадачи като *(i)* определяне границите на контекста и *(ii)* определяне на класа на извлеченото понятие [44]. Други групи обучават CRF модели с текстови характеристики, които са разширени с изхода от система, базирана на правила за разпознаване на именовани единици [20]. Други като [25], [27] използват CRF модели, комбинирани със съществуващи системи за разпознаване на именовани единици и чънкери или подобни алгоритми, базирани на изхода от ресурси, обогатени със знание за света. Има и подходи, където се прилагат характеристики от дистрибутивна семантика за обучение на CRF модели с ограничено участие на учител (semi-supervised). Най-добрата система от състезанието от 2010 [12] постига 0.852 $F-$*мярка*. Тя използва алгоритъм, подобен на CRF - дискриминативен

[3]i2b2 (Informatics for Integrating Biology and the Bedside) - Национален център за биомедицински компютинг, САЩ. https://www.i2b2.org/

[4]https://www.i2b2.org/NLP/HeartDisease/PreviousChallenges.php

полу-Марков HMM, обучен, използвайки пасивно-агресивно онлайн обновяване (passive-aggressive online update). Авторите извличат голям брой характеристики, между които са низовете на думите, конктекст, изречение, секция, характеристики на документа, характеристики извлечени от външни програми като cTakes[5] [47], MetaMap[6] [2] [3], ConText[7] [23], статистически парсери и др.

Ние участвахме в състезанието на i2b2 през 2012 г. Нашият подход включва комбинация от подходи за машинно самообучение и такива, базирани на правила. За извличането на събития са използвани техники за машинно самообучение, комбинирани с резултати от MetaMap и последваща обработка, базирана на правила. За извличането на времеви изрази TIMEX3 се използва голямо множество от регулярни изрази. Тъй като само работата по извличане на събития е дело на автора, в следващите секции ще разглеждаме само методите и оценката за извличане на събития и ще обсъдим интеграцията им с модулите за извличане на TIMEX3-маркери.

3.2.2 Ресурси

Това изследване е проведено над данни, предоставени от Състезанието i2b2 2012[8], Задача за извличане на темпорални релации. Документите са ПЗ в свободен текст на английски език на пациенти, диагностицирани с различни болести. Учебните данни са в XML-формат, съдържат оригиналния текст и 4 типа анотации (под формата на XML маркери) – EVENT, TIMEX3, TLINK и SECTIME. Организаторите предоставят детайлно описание на всеки тип анотация - структурата, покритието и съответни атрибути. Записите са организирани в 3 секции: административни данни – (i) дата на приемане и на изписване; (ii) история на заболяването и (iii) курс на лечение. Секциите са сравнително добре сегментирани. Повечето от документите съдържат около 40-60 изречения. Таблица 3.8 представя статистически данни на корпуса.Анотираните уюебни данни бяха предоставени 6 седмици преди крайния срок за предаване на готова система.

[5] cTakes - http://ctakes.apache.org/

[6] MetaMap - http://metamap.nlm.nih.gov/

[7] ConText algorithm - https://code.google.com/p/negex/

[8] Извличане на темпорални релации i2b2 2012 - https://www.i2b2.org/NLP/TemporalRelations/

данни	документи	думи	изречения	събития	уникални събития	TIMEX3 маркери	уникални TIMEX3
учебни	187	96 053	11 543	16 468	9 129	2 366	1 598
тестови	120	80 429	9 339	13 594	7 873	1 820	1 223

Таблица 3.8: Описателна статистика за учебните и тестови данни

3.2.3 Метод

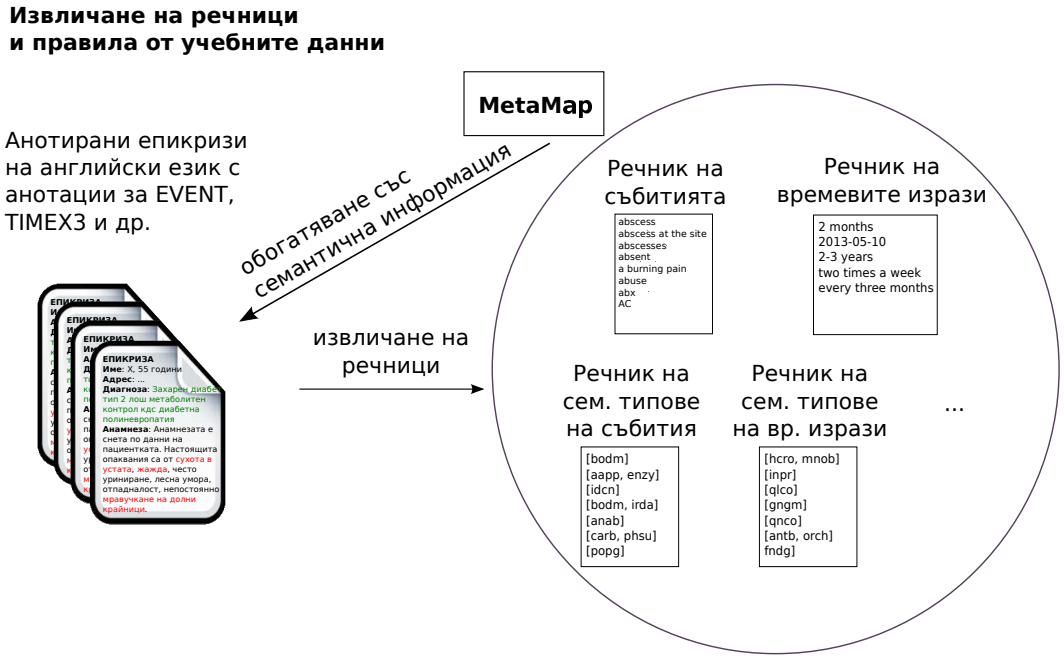
Решаваме задачата за разпознаване на събития чрез методи за машинно самообучение с учител. На последваща фаза прилагаме обработка базирана на правила за прецизиране на резултатите и за уточняване на атрибутите на всяко събитие. Обща схема на обработката е представена на Фигура 3.4 и Фигура 3.5.

Предварителна обработка

На първата фаза правим предварителна обработка на учебните данни. Подготвяме речници от думите, които участват в анотираните изрази. Всички фрази, които са маркирани като събития, са извлечени в речниците. Дублиращите са премахнати и така оформилият се списък наричаме *речник от събития*. Процедурата е повторена за анотации от тип TIMEX3 и така изготвяме *речник от времеви изрази*.

На втората стъпка обработваме всички текстове от учебното множество с MetaMap [2]. В резултат получаваме списък от срещания на медицински понятия в текстовете. MetaMap извлича съвпадения на фрази от текста с медицински понятия, които са налични в UMLS, съответното предпочитано понятие в UMLS, семантичния му тип и флаг за отрицание, ако срещането в текста се намира под въздействието на отрицание. В това изследване ползваме MetaMap2011v2.

На третата стъпка от предварителната обработка учебните данни се обработват последователно от поредица от програми за лингвистичен анализ в платформата GATE [11]. Интегрираните модули ANNIE [10] се ползват за отделяне на словоформи и думи и сегментиране по изречения. След това се прилагат



Фигура 3.4: Извличане на речници на събития, темпорални изрази, съответни семантични типове и други от обогатените с лингвистична и семантична информация епикризи.

анализатор на части на речта и чънкер от пакета OpenNLP [9]. Освен това се използва Snowball stemmer [57] [41] [51] за определяне на корените на думите.

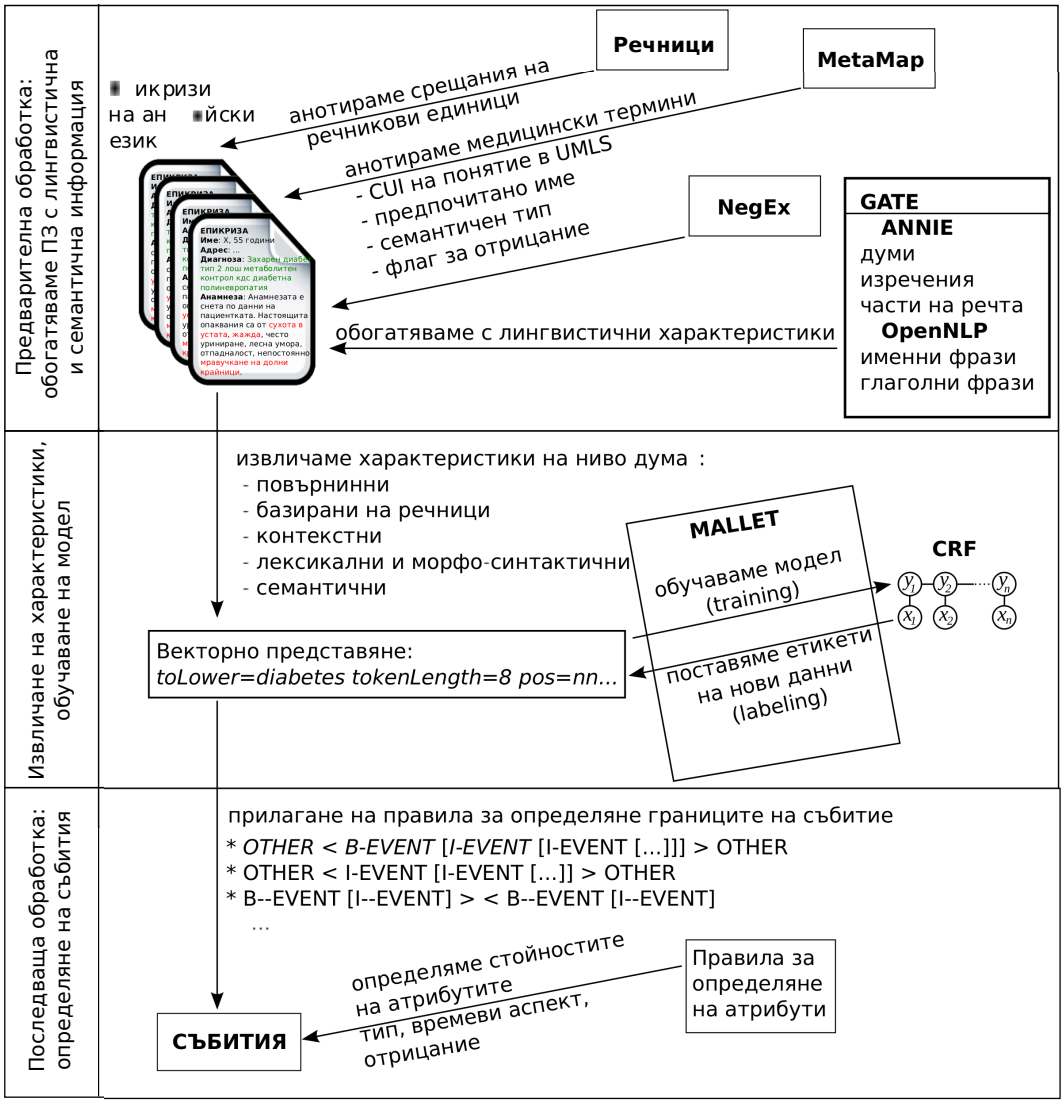
Подбиране на характеристики за машинно самообучение

Моделът за разпознаване на събития е обучен с алгоритъм CRF, използвайки следните 5 групи характеристики.

1. повърхнинни
2. базирани на речници
3. контекстни
4. лексикални и морфо-синтактични
5. семантични

Тези характеристики са извлечени за всяка словоформа от учебните документи. Пунктуацията е изключена.

[9] OpenNLP - https://opennlp.apache.org/



Фигура 3.5: Обща схема на извличането на събития от епикризи.

Повърхнинни характеристики

Първата група от характеристики са *повърхнинни характеристики*. Те са извлечени чрез прилагането на шаблони и прости манипулации върху всяка словоформа. Такива характеристики са:

низ_на_думата (низ);

низ_на_думата_с_малки_букви (низ);

дължина_на_думата (дискретна стойност);

съдържа_ли_цифри (булева стойност);

число_ли_е (булева стойност);
съдържа_ли_само_букви_и_цифри (булева стойност) и
ортографски_запис (низ).

Пример 3.2.1. *За думата* **Discharge** *повърхнинните характеристики и техните стойности са както следва:*
TOKEN=Discharge TOKEN_LOWER=discharge LENGTH=9
HAS_DIGIT=no IS_NUMBER=no IS_ALPHANUMERIC=no
ORTHOGRAPHY=upperInitial

Характеристики, базирани на речници

Втората група от характеристики е получена чрез ползване на речниците като отправна точка. За всяка дума добавяме по един атрибут, съответстващ на наличието ѝ в речниците: (*i*) дали се съдържа като термин в речниците, (*ii*) дали е част от термин в речниците или (*iii*) не се съдържа въобще. Въпреки че на тази стъпка целта ни е да разпознаем само събития, ние ползваме и речника с TIMEX3-изрази, които извлякохме на предварителната фаза. Причината да включим речника с времеви изрази TIMEX3 е, че тази характеристика може да се окаже дискриминативна, тъй като времевите изрази и събитията са непресичащи се множества. Така ако една дума е термин от речника с времеви изрази, би трябвало алгоритъмът да не се опитва да я класифицира като събитие. За всеки речник дефинираме три номинални стойности за тази характеристика:

IS_TERM – ако речникът съдържа запис, еквивалентен на входната дума;

IS_PART_OF_TERM – ако входната дума не е термин в речника, но се съдържа в някой от записите му;

NO – ако входната дума не е еквивалентна на никой от записите и не се съдържа в никой от записите.

Пример 3.2.2. *За думата* **Discharge** *характеристиките, базирани на речници и техните стойности са както следва:*
EVENT_GAZE=isEvent TIMEX3_GAZE=no

Контекстни характеристики

Третата група от характеристики са контекстните. Ние ползваме биграми отляво и отдясно на входната дума.

Пример 3.2.3. *За думата* **Discharge** *във фразата "Discharge date:" контекстните характеристики са както следва:*
L_BIGR= R_BIGR=date

Лексикални и морфо-синтактични характеристики

Четвъртата група са лексикалните и морфо-синтактични характеристики. Те са извлечени с последователност от модули в платформата GATE [11] и включват:

част_на_речта (номинална стойност) - за всяка дума;

маркер_за_чънк (номинална стойност) - маркер, поставен от чънкера (*B-NP* - начална дума на групата на съществителното, *I-NP* - дума от групата на съществителното, която не е първа, *B-VP* - начална дума на групата на глагола и др.);

корен (низ) - корен на входна дума.

Пример 3.2.4. *За думата* **Discharge** *във фразата "Discharge date:" лексикалните и морфо-синтактични характеристики са както следва:*
POS=NNP CHUNK=I-NP STEM=discharg

Семантични характеристики

Петият тип характеристики са семантичните. Това са семантичните типове, приложени към медицинските понятия, разпознати от MetaMap. Резултатите, получени от MetaMap, се запазват като MetaMap анотации. Всяка една от тези анотации има следните характеристики: *текст на съвпадението*; съответстващ *етикет на понятие*; *флаг за отрицание* и *семантични типове*. Ако дадена дума се покрива от MetaMap анотация, семантичният тип (или обединението на семантичните типове) на анотацията се копират в характеристиката *SEMANTIC_TYPE* на тази дума. Когато не е извлечен семантичен тип

за конкретната дума, това е индикация, че тази дума най-вероятно не е част от медицинско понятие (събитие или времеви израз). Събитията в тренировъчните данни често съдържат медицински понятия, разпознати от MetaMap, а времевите изрази също се анотират в известна степен от MetaMap.

Пример 3.2.5. *За думата* **Discharge** *семантичната характеристика е една: SEMANTIC_ TYPE=[clna]*

Обучение и поставяне на етикети

Моделите за извличане на събития са обучени с алгоритъм CRF [31] за разпознаване на последователности, наличен в MALLET [10] [34] и са ползвани параметри по подразбиране. Експериментирахме с няколко различни множества атрибути, комбиниращи характеристики от описаните горе. В Секция 3.2.4 са обяснени специфичните за всяка система характеристики.

Анотираме векторите съответстващи на всяка дума с класове, ползвайки BIO-анотация (Begin/Inside/Outside) както е показано тук:
B-EVENT – първа дума на събитие;
I-EVENT – дума която е част от събитие и се намира след началото;
OTHER – дума която не е подниз от събитие.

Правим 10-fold cross-validation за проверка на резултатите – моделите се обучават с 90% от учебните данни и се тестват на останалите 10%. Успешните модели са повторно обучени над цялото множество от учебни документи и са тествани върху тестовите данни. Резултатите са описани в Секция 3.2.4.

Последваща обработка

Моделът, който сме обучили, определя броя от думите, които попадат в едно събитие. Последователността от думи се счита за събитие, ако съответните им етикети/класове, добавени от модела, са както следва:

OTHER < B-EVENT [I-EVENT [I-EVENT [...]]] > OTHER
or *OTHER < I-EVENT [I-EVENT [...]] > OTHER*

[10] MALLET - http://mallet.cs.umass.edu/

където не е задължително елементите в счупени скоби да присъстват и всяка една такава последователност е предхождана от негативни примери.

В изречението: “She presented with left upper quadrant pain as well as nausea and vomiting which is a long-standing complaint.” последователността “*left upper quadrant pain*” е определена като събитие, защото думите в изречението са маркирани от модела, както е показано на Таблица 3.9:

Маркиране/разпознаване на събития със *Система 2*	
She	OTHER
<начало събитие 1>	
presented	**B-EVENT**
<край събитие 1>	
with	OTHER
<начало събитие 2>	
left	**B-EVENT**
upper	**I-EVENT**
quadrant	**I-EVENT**
pain	**I-EVENT**
<край събитие 2>	
as	OTHER
well	OTHER
as	OTHER
...	...

Таблица 3.9: Маркиране/разпознаване на събития със *Система 2*

Модалността, отрицанието (позитивно/отречено) и типа на събитията се уточняват чрез прилагане на множество от правила и ползване на пакета NegEx software[11] [8].

Модалността се определя чрез анализиране на учебните данни за думи, сигнализиращи типовете модалност. Атрибутът модалност може да приема 4 стойности: FACTUAL, CONDITIONAL, POSSIBLE, and PROPOSED. Събития-

[11] NegEx - https://code.google.com/p/negex/

та с модалност FACTUAL доминират в записите. Те са 96% от всички събития, а другите събития са по около 1%–1.5% за всяка стойност. По тази причина стойността FACTUAL приемаме за стойност по подразбиране, а за останалите категории прилагаме разпознаване, базирано на ключови думи от речници. Създаваме по един речник със "сигнални" думи, като анализираме контекста на събитията в учебните данни. Например, речникът със "сигнални" думи за модалност POSSIBLE и техните съответни честоти на срещане в учебния корпус са както следва: *question* (24), *possible* (23), *likely* (21), *most likely* (15), *possibly* (9), *presumed* (6), *questionable* (5), *probable* (2), *suspected* (2), *apparent* (1). Тези речници подаваме на софтуера NegEx вместо списъци от термини за отрицание. Пускаме NegEx без да променяма правилата му, т.е. използвайки правилата за отрицание. Дори и този подход да е по-скоро наивен, резултатите са със сравнително висока точност.

За определяне на отрицанието (дали събитието има позитивен смисъл или е отречено), системата използва резултатите от NegEx. Списъкът със "сигнални" думи в NegEx е обогатен с допълнителни, специфични за учебния корпус изрази.

Типът на събитието се определя на базата на филтри за семантичния тип и речници, извлечени от учебния корпус. Филтрите са създадени след анализ на анотациите, направени от MetaMap. Взети са предвид семантични типове, определени за всяка от разглежданите анотации и зависимостите им с типовете събития, определени от задачата на i2b2. Типът на едно събитие зависи от семантичния тип на думите, които го съставят. В допълнение към списъците създадени от учебни данни, използваме и списък от лекарствата в ICD-9 за разпознаване на събития от тип TREATMENT.

3.2.4 Експерименти и резултати

Резултати при предаване на системата

Задачата по извличане на събития решаваме с комбинация от подходи за машинно самообучение и подходи, базирани на правила. Модел, обучен чрез машинно самообучение с учител, поставя етикети върху всяка дума, но решението за това, кои етикети формират едно събитие и кои ще са неговите атрибути,

се взема на базата на правила (както е обяснено в Секция 3.2.3).

Базовият модел за това изследване е обучен върху ограничено множество от характеристики: контекстни характеристики (лява и дясна биграма), повърхностни характеристики и характеристики от речници. Тази система тук се споменава като *Система I*. *Система II* е моделът, който ползвахме за предаване на нашата система на състезанието. В допълнение към характеристиките, използвани в *Система I*, *Система II* включва семантични типове, извлечени от MetaMap. Резултатите са показани на Таблица 3.11 заедно с резултатите от другите системи. Резултатите, които са показани в тази таблица, са генерирани от програма за оценяване на задачата по извличане на събития, предоставена от организаторите.

Резултати и експерименти след предаване на системата След предаване на *Система II* на организаторите на състезанието, продължихме да работим и направихме подобрения в разпознаването на събития. Сравнение между отделните системи е показано на Таблица 3.11.

За удобство тук ще разгледаме характеристиките на базовата система, системата, чиите резултати предадохме на организаторите, и 2 системи, които усъвършенствахме след това.

Система I: Базова система. Включва повърхностни характеристики, речници и биграми.

Система II: Система, която предадохме за участие в състезанието. Тя включва характеристиките от *Система I* и характеристики на понятията, извлечени чрез MetaMap.

Система III: Тази система включва характеристиките от *Система II* и лексикални и морфо-синтактични характеристики като части на речта, чънкове, ортографичен запис, стем.

Система IV: Включва характеристиките на *Система III* и модификация в процедурите за последваща обработка. В *Система IV* етикетите се поставят от същия модел, както и в *Система III*, но процедурата за подбор на съответна последователност от думи, които представляват събитие, е различна. В Системи I-III подборът се прави, както е описано в Секция 3.2.3; в *Система IV* вземаме предвид, че няколко събития могат да бъдат последователни (както е показано в Пример 3.2.6).

Пример 3.2.6. *< B–EVENT [I–EVENT] > < B–EVENT [I–EVENT] >*
В Система IV и двете последователности ще бъдат извлечени като събития. Така в изречението "Her pain resolved after surgery and she has been doing well since, although at baseline now she is minimally ambulatory from bed to commode.", и двата последователни израза "Her pain" и "resolved" се приемат за две отделни събития, защото са маркирани, както е показано на Таблица 3.10.

Маркиране/разпознаване на събития със *Система IV*	
<начало събитие 1>	
Her	**B-EVENT**
pain	**I-EVENT**
<край събитие 1>	
<начало събитие 2>	
resolved	**B-EVENT**
<край събитие 2>	
after	OTHER
...	...

Таблица 3.10: Маркиране/разпознаване на събития със *Система IV*.

Резултатите от четирите системи са описани в Таблица 3.11.

Сравнявайки четирите системи, виждаме, че относителният брой на разпознати събития се увеличава със всяка следваща система и множество от характеристики. От 69.52% в системата, предадена в състезанието, до 83.26% в последната и най-добра *Система IV*. Има също и значително покачване на покриваемостта от 65% в началото, до 76% в *Система IV* (с 11 точки). В замяна на това се понижава леко точността с 2 точки. Хармоничното средно $F-$*мярка* се увеличава с около 6 точки и достига 82%.

Съпоставяме резултатите от *Система IV* с резултатите, публикувани в [53]. Според това оценяване нашите резултати са сравними с 9-то място по $F-$*мярка* от всички системи участвали в задачата.За жалост не успяхме да предадем *Система IV* в искания срок

Система	*Система I*	*Система II*	*Система III*	*Система IV*
Златен стандарт	13 594	13 594	13 594	13 594
Анотирани събития	9 450	10 917	10 959	11 319
Дял анот. събития	(69,52%)	(80,31%)	(80,62%)	(83,26%)
Покриваемост	0,651020	0,718686	0,735365	0,757872
Точност	0,925303	0,885493	0,909995	0,908312
Средно точн.&покр.	0,760393	0,788838	0,808540	0,821267
F−мярка	0,760382	0,788813	0,808490	0,821223
Модалност	0,933152	0,931658	0,930180	0,932005
Отрицание	0,954637	0,950857	0,951373	0,954857
Тип	0,712071	0,705148	0,705877	0,702273

Таблица 3.11: Оценяване на извличането на събития върху тестовото множество от данни с програми предоставени от i2b2.

3.2.5 Дискусия и анализ на грешките

Нашите резултати показват, че хибридна система, разчитаща на машинно самообучение и правила, може да достигне сравнително висока *F−мярка* при разпознаване на събития, вероятно близка до съгласуваността между хората, анотирали ръчно златния стандарт, и това може да бъде достигнато с използване на малко на брой характеристики на текста. MetaMap и NegEx бяха единствените семантични екстрактори, които използвахме. С това искаме да подчертаем, че те са богат източник на семантични характеристики. Освен това те са публично достъпни, но работят само за английски език.

Приноси и анализ на грешки на MetaMap

MetaMap и семантичните типове, които програмата прилага към всяко анализирано понятие, имат централна роля в генерирането на характеристики на събитията, обучението на моделите и разпознаването на събития. Тук представяме някои детайли от приноса на характеристиките, извлечени от MetaMap. При оценяване на грешките ръчно съпоставихме анотациите на събития от учебните данни с тези на извлечените понятия от MetaMap. Забелязахме, че 12 566 понятия, извлечени от MetaMap, не съвпадат с нито едно събитие от учебните

данни (това е около 76% свръх генериране (overgeneration)) и само 158 събития от корпуса нямат съответно понятие в MetaMap (0.95%). В много случаи анотациите от MetaMap покриват изрази, съответстващи на пациент като *male, female, patient* и *Mr.* Тъй като те никога не са маркирани като събития в учебния корпус, тези анотации не пречат на обучението на моделите. От друга страна, често датите в текста не са разпознати правилно от MetaMap и им е присвоен грешен семантичен тип. Този факт влияе върху разпознаването на събития, защото семантичният тип *Temporal concept* е знак за темпорален израз и неявен знак, че дадената дума не е събитие. В тези случаи, грешните семантични типове внасят шум в описанието на негативните примери. Въпреки тези недостатъци, резултатите в Таблица 3.11 показват, че семантичният тип допринася за повишаване както на точността, така и на покриваемостта при разпознаване на събития. Това се вижда от сравнението със системата, базирана само на повърностни характеристики (*Система I*).

3.2.6 Резюме

Резултатите, които демонстрирахме тук, показват, че дори системи, работещи със сравнително базови характеристики, могат да постигнат разумни резултати. Те също показват, че анотирането на златния стандарт е било направено добре и че данните са надеждни. Показват, че темпоралността в клиничните текстове може да бъде наблюдавана по автоматичен начин като задача в обработката на естествен език в биомедицински текстове. Тези резултати предполагат също така значителна възможност за подобрение. Ако организаторите продължават да повтарят това състезание с повече данни, ще дадат възможност да се доразвият създадените системи и да се постигнат още по-високи резултати, така че да се приближат до софтуер за използване в реалния свят.

Приложените тук методи могат да се приложат със същия успех и върху данни на български език, ако са налични етикети за понятията в семантични ресурси в UMLS на български. В този смисъл работата по откриване на събития в пациентски записи на английски език е и стъпка напред към автоматичното структуриране на ПЗ на български.

Глава 4

Интеграция в прототипи

4.1 Вграждане на МС за класификация на атрибути в среда за автоматично откриване на потенциални диабетици в хранилище с големи данни

4.1.1 Въведение

Методите, които разработваме за извличане на информация от медицински документи, интегрираме успешно и в система за анализ на амбулаторни документи. Системата работи с милиони амбулаторни листи на пациенти на български език и методите за ОЕЕ се прилагат върху специфични секции от записите. Компонентите за извличане на информация са разработени и тествани в продължение на няколко години и в различни проекти, за да постигнем висока точност и покриваемост на резултатите. Системата, която представяме, включва модули за "business intelligence" и такива за ОЕЕ. ОЕЕ модулите се ползват за нормализация, изчистване на данните, извличане на знание от голям обем данни. Комбинацията от тези инструменти спомага да създадем Регистър на на диабета в България, като ни дава възможност да открием потенциални диабетици, които не са формално диагностицирани с диабет. Приносът на автора на дисертацията е в разпознаването на потенциални диабетици чрез анализ на свободен текст.

Целта на това изследване е да разпознаем пациенти, които имат диабет, но не са формално диагностицирани с тази болест и тази диагноза не присъства в амбулаторните им документи. Като сигнали за наличието на диабет считаме: (*i*) високи стойности на *кръвна захар* и *гликиран хемоглобин* в текста на секцията *Лабораторни резултати*; (*ii*) споменаване на определени лекарства в секция *Лечение* – ако пациентът има диабет, той/тя би приемал/а и подходящи медикаменти, и (*iii*) факти, коментари или описания в анамнезата, които сочат, че пациентът може би има диабет и/или усложнения на диабета. Анализираме съответните секции от документа, като прилагаме OEE, ползваме специфични характеристики, извлечени от BITool[1]. Медицинските експерти поставят като критерий за случване на събитието "има диабет" наличието на някой от следните факти: *висока кръвна захар*; *висок glycated хемоглобин* в текста на секцията *Лабораторни резултати*; приемането на лекарства, които са подходящи за лекуване на диабет, споменати в секцията *Анамнеза*; други факти, споменати в анамнезата, описващи диабет или неговите симптоми/усложнения. За извличането на тези събития направихме конкордансер и резултатите от него са обект на по-нататъчна обработка с OEE модули, за да се потвърди или отхвърли хипотезата "има диабет".

4.1.2 Дефиниция на задачата

За да се потвърди/отхвърли хипотезата "има диабет", прилагаме методи за машинно самообучение с учител, за филтриране на текстовите чънкове от секциите със свободен текст в амбулаторните документи. За момента експериментираме с конкорданс на думата "*диабет*". Подобен анализ може да се направи и за друга болест или симптом, който ни интересува. Избираме документи, в които няма експлицитна диагноза *диабет*. Искаме да покажем, че OEE може да помогне да намерим противоречиви данни в хранилището за документи (по този начин, ще подпомогнем и почистване на данните) и/или да разширим метаданните за конкретния запис, ако намерим фрагменти в текста, които потвърждават диагнозата "диабет".

[1]BITool - Business Intelligence Tool, софтуер за управление и анализ на големи бази данни

4.1.3 Входни данни

Входните данни са текстови фрагменти около думата "*диабет*", извлечени от конкордансер. Извлечени са 67 904 различни фрагмента, които идват от 156 310 пациентски записа на български език, в които не се среща диагнозата "*диабет*". Всеки фрагмент съдържа низа "*диабет*" и думите, които се срещат в прозорец от 6 низа отляво и отдясно в контекста на думата. Текстът е сведен до корени на думи. Нашата задача е да класифицираме тези фрагменти, като потвърдим или отхвърлим хипотезата "има диабет".

Пример 4.1.1. *Това са няколко фрагмента, които демонстрират разнообразието от позитивни и негативни примери спрямо нашата хипотеза.*
*(i) NEG Фамилност- обременен/а-***диабет***ици по майчина линия/.*
(ii) NEG Необходимо е изключване на стероиден **диабет***; насочва се към ТЕЛК...*
(iii) POS Покачва кръвно налягане. Има **диабет***. Оплаква се от сърцебиене...*

Примери (*i*) и (*ii*) са негативни - първият говори за наследственост, но не потвърждава диагнозата *диабет*, а във втория диагнозата е отречена. Пример (*iii*) е позитивен. Често фактите, потвърждаващи диабет,са изказани във фрази с дължина около 3-4 думи.

4.1.4 Експерименти и резултати

За разрешаване на задачата, прилагаме комбиниран подход, базиран на правила и на машинно самообучение: (*i*) итеративно филтриране, базирано на правила, за да се редуцира броят на записите, които потенциално биха потвърдили нашата хипотеза; (*ii*) трениране на модел чрез машинно самообучение за автоматично класифициране на записите от редуцираното множество на позитивни и негативни спрямо хипотезата.

Грубо филтриране, базирано на правила

Тъй като данните са извлечени от документи, в които диагнозата "диабет" не е спомената явно, очакването е, че по-голямата част от извлечените фрагменти ще бъдат негативни спрямо хипотезата. На първата стъпка решихме да

филтрираме колкото можем повече от негативните примери с прилагането на правила и така да редуцираме размера на входните данни, които ще класифицираме. Няколко шаблона за негативни примери се срещат сравнително често в нашите данни - напр. “няма данни за диабет”, “няма диабет в семейството” и др., които са в около 10% от нашите записи. Това са изрази, които говорят за фамилна обремененост или отричане на диагнозата “диабет”. Итеративно създаваме множество от регулярни изрази на ниво корен на думата, които да съвпадат с такива фрагменти. С множество от 41 такива изрази, използвани като филтър, броя на фрагментите в нашето входно множество данни се намали до 26 000 (което е около 1/3 от началния брой).

Класификация на позитивни/негативни примери с МС

На втората фаза извлякохме на случаен принцип от редуцираното множество записи две подмножества от записи - едно с 282 фрагмента и едно с 1 000 фрагмента и ги анотирахме. Първото подмножество беше използвано за анализ (development set), използвахме го за подбор на характеристики и за начални тестове. То съдържа 74 позитивни и 208 негативни примера, докато второто съдържа 187 позитивни и 813 негативни примера. Използвайки различни характеристики на тези записи и алгоритми за класификация, ние проверихме доколко е приложимо машинното самообучение за автоматично извличане на записи, отнасящи се до “има диабет” (подобно на [23], [39], [46]) и опитахме да поставим база за сравнение за тази задача. Различното в нашия подход, спрямо изброените е, че за тази задача ние моделираме реални данни в голям обем.

Във фазата на предварителна обработка свеждаме думите от текстовите фрагменти, до корени за да преодолеем морфологичната инфлексия. Поради големия брой абревиатури, дозировки, лабораторни резултати и др., които съдържат пунктуационни знаци, автоматичното разделяне на текста на изречения с наличните инструменти не е достатъчно надеждно и затова не го прилагаме. Постигайки сравнително високи резултати, ползвайки само повърхнинни и структурни характеристики, ние доказваме приложимостта на подхода. Експериментирахме с 2 вида векторни характеристики - булеви и номинални. Векторите, характеризиращи всеки фрагмент, който искаме да класифицираме, съответстват на предварително дефинирани корени, биграми и триграми, които

са предефинирани. Когато ползваме булеви стойности на векторите, всяка характеристика във вектора отговаря на съответен корен, биграма или триграма и е 1, ако съответната характеристика е налична в текстовия фрагмент или е 0, ако отсъства.

Списъка от характеристики съставихме след анализиране на множеството от корени на думи, които се срещат в данните. Направихме експерименти с 3 различни множества от характеристики, които съответстват на експериментите, описани по-долу. Опитахме няколко алгоритъма върху едно и също множество от характеристики: Naïve Bayes, J48, SMO и JRip, всички с булеви характеристики: JRip и J48 се справиха най-добре. Направихме също така и класификация с номинални характеристики, ползвайки алгоритъма MaxEnt, и той показа по-добри резултати за точност от останалите в смисъла на точност. Използвахме същите характеристики - корените на думите, биграми и триграми. В това изследване се интересуваме най-вече от точността върху позитивните примери, защото по този начин можем да добавяме нови пациенти към регистъра с минимални усилия и голяма сигурност съгласно разпознатите позитивни примери като пациенти, които имат диабет, но не са били формално диагностицирани. Чрез прилагане на няколко филтъра с висока точност, като тези, които вече описахме накрая можем да получим и добра покриваемост. Следват детайли за експериментите. Резултатите са описани в Таблици 4.1, 4.2 и 4.3.

Експеримент 1

Използваме 93 характеристики, които съответстват на корените на термините, срещани в текста на позитивните примери (изключваме числата). Резултатите от 10-fold cross-validation с алгоритмите, които се справиха най-добре с класификационната задача,са показани в колони 2–7 на Таблица 4.1. J48 разпозна само 63.1 % от позитивните примери, но точността, постигната от JRip, е окуражаваща - 91.2%. Правилата, извлечени от JRip, са само 2, но очевидно те подхождат достатъчно добре на данните

Rule 1: (family = false) and (sugar = true) and (noninsulindependent= true)
=> class=pos (30.0/2.0)

Rule 2: (family = false) and (sugar = true) and (treat = true)

=> class=pos (6.0/0.0)

При класификация на по-голям обем данни, тези 2 правила може да се окажат недостатъчни и затова продължихме да работим върху подбора на характеристики.

Експеримент 2

В този експеримент използваме 112 текстови характеристики, които съответстват на корените на термините,срещани в позитивните и негативните примери. Термините са филтрирани предварително от експерт. И двата алгоритъма - JRip и J48, изведоха по-ниски резултати от Експеримент 1 и постигнаха точност под 70% (Таблица 4.1, Експ. 2).

Експеримент 3

В този експеримент ползвахме алгоритмите за автоматичен подбор на характеристики за машинно самообучение. Началното ни множество съдържаше 10 576 атрибута, съответстващи на корените на всички думи в множеството за анализ. Приложихме chi-squared алгоритъм за оценяване на атрибути, наличен във Weka [22], за да определим представителността на атрибутите спрямо множеството от текстови фрагменти. В резултат получихме 151 атрибута. В експерименти 3a и 3b използвахме тези атрибути в комбинация с биграми и триграми (Таблица 4.2). Резултатите от двата алгоритъма бяха по-добри при добавянето на биграми и с добавянето на триграми се подобриха още повече. JRip достигна 65.3% точност и 73.1% когато добавихме триграми като атрибути. J48 постигна точност от 86.1% след добавяне на биграми и 87.2% с добавяне на триграми към множеството от атрибути. В дървото, построено от J48, може лесно да се види важността на биграмите и триграмите като дискриминативни характеристики - от 17 листа на дървото, само 4 са униграми; останалите са биграми и триграми. Обясняваме си това с факта, че триграмите съдържат реда на корените и дори в някои случаи представляват конкретни изрази, означаващи наличието/отсъствието на "сигнализиращите изрази", които потвърждават/отричат нашата хипотеза. Такива са: "инсулинозависим захарен диабет", "неинсулинозависим захарен диабет", "със зах диабет" , "майка с диабет". Според нашите наблюдения последният израз сигнализира отричане на диабет в конкретния текстов фрагмент, защото това е фамилна анамнеза и нейната дъл-

жина обикновено покрива достатъчно голяма част от фрагмента, така че не може да се помести и друго описание, което да се отнася до пациента и да потвърди хипотезата. Резултатите показват, че до този момент добавянето на още характеристики води до по-добра точност (с изключение на JRip в експеримент 1) и също така покачваща се покриваемост с JRip.

	"Експ. 1: 93 pos features; само корени; 10-fold cross-validation"						"Експ. 2: 112 pos/neg features; само корени; 10-fold cross-validation"					
	JRip			J48			JRip			J48		
Клас	P	R	F	P	R	F	P	R	F	P	R	F
Поз.	91.2	16.6	28.1	66.7	21.4	32.4	61.1	29.4	39.7	65.8	52.4	58.3
Отриц.	83.9	99.6	91.1	84.4	97.5	90.5	85.5	95.7	90.3	89.5	93.7	91.6
Средно ар.	85.2	84.1	84.1	81.1	83.3	79.6	80.9	83.3	80.8	85.1	86	85.4

Таблица 4.1: Резултати от Експеримент 1 и 2 с ръчно подбрани характеристики.

	"Експ. 3a: 151 автоматично подбрани атр.: корени + бигр.; 10-fold cross-validation"						"Експ. 3b: 151 автоматично подбрани атр.: корени + бигр. + тригр.; 10-fold cross-validation"					
	JRip			J48			JRip			J48		
Клас	P	R	F	P	R	F	P	R	F	P	R	F
Поз.	65.3	41.2	50.5	86.1	36.4	51.1	73.1	40.6	52.2	87.2	36.4	51.3
Отриц.	87.5	95	91.1	87.1	98.6	92.5	87.6	96.6	91.9	87.1	98.8	92.6
Средно ар.	83.4	84.9	83.5	86.9	87	84.8	84.9	86.1	84.5	87.1	87.1	84.9

Таблица 4.2: Резултати от експерименти 3a и 3b - автоматичен подбор на характеристики, ползване на биграми и триграми.

Експеримент 4

Обучихме модел с алгоритъма MaxEnt, използвайки всички текстови характеристики, включвайки биграми (Експ. 4a) и биграми и триграми (Exp. 4b) като номинални стойности. Всички думи бяха първо сведени до корени. Тези два експеримента произведоха много подобни резултати и се справиха по добре от J48 и JRip по отношение на точността. Най-високата достигната точност на позитивните примери е 91.5%, когато включихме биграми и триграми. Ползването на MaxEnt с номинални характеристики има предимството, че наборът от

	"Експ. 4a: вс. корени + биграми; 10-fold cross validation"			"Експ. 4b: вс. корени + биграми + три; 10-fold cross validation"		
	MaxEnt					
Клас	P	R	F	P	R	F
Позит.	91.3	22.6	36.2	91.5	20	32.8
Отриц.	85.6	84.2	85	85.6	84.2	84.9
Средно ар.	88.45	53.4	60.6	88.55	52.1	58.9

Таблица 4.3: Експерименти 4a и 4b с MaxEnt и всички текстови характеристики, биграми и триграми

характеристики не е предварително зададен. Подобните резултати подсказват, че използване на MaxEnt с тези характеристики може да служи като база за сравнение за бъдещи изследвания. Позитивните примери в нашите данни са толкова рядко срещани (защото по принцип диабетиците са формално диагностицирани), че качеството на учебните данни има съществено влияние върху обучението на модела. Обучението може да се подобри, ако си набавим още такива записи, и това ще стане, когато по-голям масив от предишни години станат налични. Въпреки това текущите резултати показват, че хибриден метод, комбиниращ правила и машинно самообучение, може да се използва за доказване на хипотезата “има диабет” с голяма точност.

4.1.5 Резюме

Този резултат представя първи стъпки към създаването на компютърна платформа за работа с големи масиви от данни в областта на биомедицината с приложение от реалния свят. Очевидно е, че решения за медицински случаи не могат да се вземат абсолютно автоматично и затова окончателното мнение би трябвало винаги да е човешко съображение. Но в същото време автоматичната обработка на текст на пациентските записа дефинира съвсем нов хоризонт за повечето от задачите, свързани със здравните анализи.

Модулите за извличане на информация са разработени внимателно, за извличане на ограничен брой именовани единици и събития. Те са тествани в различни сценарии и тяхната успеваемост постепенно се подобрява, използвайки

хибридни подходи, базирани на правила и на машинно самообучение. Извличаме автоматично от свободния текст на записите съществени изрази/имена, свързани с лечението, като имена на лекарства, дозировки, начини на приемане,като тук показахме как класифицираме записите спрямо хипотезата “има диабет” с точност 91.5%.Предлагаме тези открития на специалистите, от които зависи да вземат решения с цел да се подобрят здравните политики и мениджмънта в българската здравна система. Ние мислим, че автоматичният анализ на голям обем медицински текстове може да бъде надеждна технология, ако входните данни са добре структурирани на зони (какъвто е случая с амбулаторните листи) и ако целите на задачата за извличане на информация са ясни и добре дефинирани в смисъла на именовани единици и релации.

Заключение

В тази дисертация представихме набор от техники за ОЕЕ, които са значими при изграждане на семантични системи, обработващи текст в областта на биомедицината. Предложените подходи и експерименти са свързани с два от основните компоненти на ССТ, а именно модел на предметната област и компонент за извличане на информация. Демонстрираните подходите за ОЕЕ върху текст на български език, са едни от първите опити да се структурира медицински текст на български.

Изследвахме наличността на ресурси на български език, които описват предметната област и са подходящи за вграждане в интегрирани ССТ. Липсата на изчерпателни терминологични банки, речници със семантична информация и онтологии с етикети на български език ни подтикна да изследваме приложимостта на чуждоезикови ресурси за създаване на декларативен модел на областта на български език. В резултат разработихме технология за разпознаване на парафази на термини от свободния текст на ПЗ като разглеждаме тяхни преводни съответствия в UMLS и прилагаме правила за филтриране на подходящите кандидати за парафрази/синоними.

За построяването на декларативен модел на областта предложихме също и техники за обогатяването на модела с релации между термините. Извличаме ги автоматично от кратки дефиниции в номенклатурите на UMLS. Фокусът беше над йерархичната релация **IS-A** и релацията **AFFECTS**, която свързва заболяване и органа, в който то причинава патологични изменения. Подходът е базиран на правила, комбиниращи повърхнинни, лингвистични и семантични характеристики на думите в текста. Тъй като релацията **IS-A** до голяма степен се разглежда при построяването на терминологичните ресурси и обикновено е дефинирана, особено ценна е релацията **AFFECTS**. Наличието на модели

с такъв вид връзки между понятията помага за по-прецизното извличане на информация от свободен медицински текст и структурирането му до шаблони от тип атрибут-стойност, особено при задачи за извличане на състоянието на пациента.

Създаденият модел на предметната област ползвахме при следващите изследвания за структуриране на симптоми на диабета в анамнезата на пациентските записи на български език. Разгледахме "*кръвна захар*", "*полиурично-полидипсичен синдром*" и "*промяна в теглото*", както и "*възрастта*" на пациента. Сравнихме два подходи за филтриране на изреченията, съдържащи описание на състояние - основан на правила, и на машинно самообучение. За определяне типа, атрибутите и границите на описанието на симптома в текста създадохме речници по метод отдолу-нагоре и извлякохме правила за разпознаване на симптомите в текста. С комбиниран подход, базиран на правила и на машинно самообучение демонстрирахме извличане на събития от епикризи на английски език.

Описаните по-горе техники всъщност представляват обичайният подход за обработка на текст в нова предметна област - започваме с обозиране на понятията в областта; после изброяване на важните връзки между тях и изработване на декларативен модел на областта; определяне на обекти и събития в текста, които са интересни за извличане - характеристики, атрибути; създаване на техники за извличането им. Можем да кажем, че с преминаването през всички стъпки на този процес сме навлезли дълбоко в областта на ОЕЕ в биомедицината и по-специално на ПЗ на диабетици. С влагането на всички тези елементи в цялостна система за откриване на потенциално диабетноболни пациенти чрез анализ на амбулаторни листове, показахме успешна интеграция на методите за ОЕЕ със системи за работа с голяма база данни и доказахме приложимостта и качеството на разработените компоненти.

От направените изследвания можем да кажем, че:
(*i*) извличането на релации между понятия от дефиниции на термини, които са различни от **IS-A** е сравнително по-сложен процес, но наличието на тези релации в модела е важен фактор при структуриране на медицинските описания, особено тези за статус на пациента;
(*ii*) хибридните подходи, базирани на правила и машинно самообучение, рабо-

тещи над лингвистични, повърхнинни и семантични характеристики се справят добре, дори и без наличието на външни терминологични и семантични ресурси на български език, в една сравнително добре дефинирана област, като ПЗ на диабетици, където симптомите са ясно определени и са изразени в текста;
(*iii*) разработените подходи за извличане на събития от медицински текст на английски биха били приложими за български при наличие на етикети на понятията на български в съответните терминологични и семантични ресурси;
(*iv*) за отделни задачи като класификация на изречения, съдържащи потенциални изрази, от които се интересуваме, подходите, базирани на правила и тези на машинно самообучение са взаимозаменяеми (с малък толеранс) и изборът между двата е въпрос на предпочитание и възможности за изработване на тренировъчни данни или правила за анализ;
(*v*) за преодоляване на инфлексията ползването на стемер (свеждане на думите до техните корени), се оказа подходящ подход - и да има случаи на препокриване на различни по семантика думи, те са пренебрежимо малко.
(*vi*) при класификация на текстови фрагменти с булеви вектори получаваме най-добри резултати с алгоритмите J48 и JRip, а с номинални стойности - с MaxEnt.

Насоки за бъдеща работа

ОЕЕ в биомедицината е сравнително нова област и открива нови хоризонти за повечето от задачите, свързани със здравните анализи. Бъдещите насоки в нашата работа са свързани с:
(*i*) подобряване на техниките за извличане на релации между понятията, тъй като те са от съществено значение при извличането на информация от медицински текст;
(*ii*) разработване и интеграция на нови ОЕЕ компоненти за извличане на информация от амбулаторни листове и така да се улесни създаването на регистри на пациентите от вече съществуващите записи; да се правят автоматични справки върху текстовата информация на медицинските документи; да се извличат зависимости, които могат да се наблюдават само над голям обем данни и информация за тях е налична само в текста.

Основни научни и научно-приложни приноси

Основните научни и научно-приложни приноси на дисертацията са както следва:

1. Предложен е метод за извличане на релации, различни от йерархичните *IS-A*, между медицински понятия с дефиниции в UMLS. Релациите са дефинирани в семантичната мрежа на UMLS, но не са систематично указани между понятията от интегрираните номенклатури и класификации. В метода е включено автоматично филтриране на дефинициите, които са подходящи за синтактичен и семантичен анализ с цел по-нататъшно разпознаване на концептуални релации между заболявания и засегнати от тях части и системи на човешкото тяло. На автора не са известни други подобни опити за използване на (многото различни) дефиниции на понятия в UMLS с такава цел. Макар че експериментите с прототипа за откриване на една релация се намират в начален стадий и досега е работено над ограничено подмножество "сигнализиращи" релациите думи, възможността за повторно използване на текстовете от UMLS е много важна поради обема на наличната информация в този ресурс, от една страна, и от друга – недостига на декларативни концептуални модели в медицината.
2. Създаден е прототип за извличане на релации от медицински дефиниции, с цел обогатяване на декларативен концептуален модел на предметната област. Прототипът работи над правила, базирани на повърхнинни, лингвистични и семантични характеристики.
3. Разработена е технология за разпознаване на парафази от клинични текстове на пациентски записи, като създаденият граматичен ресурс е извле-

чен от първичните документи на експерименталния корпус. Технологията е разработена над 368 записа, извлечени от 1000 епикризи на български език. За тестване са ползвани 66 записа, за които прототипът работи с точност 96% и покриваемост 89%.

4. Технологията за обработка на парафрази в клинични текстове е приложена при задача за структуриране на характеристики на пациента чрез извличането им от свободен текст и запълването им в предварително определени шаблони (в база данни). При извличане на стойности на *кръвната захар*, *промяна в теглото*, *полиурично-полидипсичен синдром*, *възраст* на пациенти-диабетици от записи на български език, както и при извличане на *събития* от медицински записи на английски език е постигната *F-мярка* както следва:
   - за *кръвната захар* - между 60% и 96,5% за отделните фази на разпознаване на симптома и атрибутите на релацията;
   - за *промяна в теглото* - между 77,8% и 93,5% за отделните фази на разпознаване на симптома и атрибутите на релацията;
   - за *полиурично-полидипсичен синдром* - 90%;
   - за *възраст* - 89,44%;
   - за *събития* от записи на английски език - 82,12%.
5. Създаден е модул за класификация на фрагменти от пациентски записи спрямо хипотезата "има диабет" с точност 91.5%, който анализира думите и низовете в 13-позиционен колокационен прозорец около думата *диабет* (6 позиции отляво и 6 позиции отдясно).
6. Извършена е прототипна интеграция на класификатора от т. 5 в компютърна платформа за работа с големи данни, предоставени от Здравната каса. Задачата на класификатора е да допринесе за разпознаване на записите на потенциални диабетици, които не са формално диагностицирани с тази болест. Крайната цел на изследванията е да се създаде Регистър на диабета в България чрез повторно използване на налични здравни записи и без допълнително утежняване на административните процедури. Има няколко начина да се разпознаят потенциални диабетици в предоставените записи (например анализ на стойности от лабораторни изследвания и приемани лекарства). Класификаторът от т. 5 е натоварен със задачата да

обработва признаци и коментари, дадени в описанията на свободен текст. Тази интеграция е първа стъпка към създаване на софтуерно приложение, което ще работи над големи данни в реалния свят.

Списък със статии по дисертацията

Основните резултати по дисертацията са публикувани в следните статии:

1. Nikolova, I. and G. Angelova. Identifying Relations between Medical Concepts by Parsing UMLS Definitions. In: Andrews, S., S. Polovina, R. Hill and B. Akhgar (Eds.), Conceptual Structures for Discovering Knowledge, Proc. of ICCS-2011, the 19th Int. Conference on Conceptual Structures, UK, July 2011, Springer, Lecture Notes in Artificial Intelligence 6828, 2011, pp. 173-186, DOI: 10.1007/978-3-642-22688-5_13
2. Nikolova, I. Paraphrase Identification in Biomedical Terminology. In: Dranidis, D., A. Kapoulas, and A. Vivas (Eds.) Proceedings of the 6th Annual South-East European Doctoral Student Conference, 19-20 September, Thessaloniki, Greece, ISBN 978-960-9416-04-7, ISSN 1791-3578, pp. 326-333.
3. Boytcheva, S., G. Angelova and I. Nikolova. Automatic Analysis of Patient History Episodes in Bulgarian Hospital Discharge Letters, In Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, April 2012, Avignon, France, pp.77-81. `http://www.aclweb.org/anthology-new/E/E12/E12-2016.pdf`
4. Nikolova, I. Unified Extraction of Health Condition Descriptions, Proceedings of the NAACL HLT 2012 Student Research Workshop, June 2012, Montreal, Canada, pp. 23-28. `http://aclweb.org/anthology-new/N/N12/N12-2005.pdf`
5. Nikolova, I., G. Angelova, D. Tcharaktchiev and S. Boytcheva. Medical Archetypes and Information Extraction Templates in Automatic Processing of Clinical Narratives, In Proceedings of ICCS 2013, 10-12 January, Mumbai, India, Conceptual Structures for STEM Research and Education, Springer, Lecture Notes in Computer Science Volume 7735, 2013, pp 106-120.
6. Boytcheva, S., G. Angelova, D. Tcharaktchiev and I. Nikolova. Extracting Patient Status Data and Temporal Event Structure from Hospital Patient Records in Bulgarian. In Proceedings of The 2'nd International Workshop on Next Generation Intelligent Medical Decision Support Systems (MedDecSup'2012), Studies in Computational Intelligence 473, Springer, pp. 203-212, 2013, ISSN: 1860_949X.

7. Temnikova, I., I. Nikolova, W. A. Baumgarther, G. Angelova, K. B. Cohen. Closure Properties of Bulgarian Clinical Text. In: Angelova, G., K, Bontcheva, and R. Mitkov (Eds.), Proc. of the Int. Conf. on Recent Advances in Natural Language Processing (RANLP 2013), September 07-13, 2013, Hissar, Bulgaria, published by Incoma Ltd., Shoumen, Bulgaria, ISSN 1313-8502, 2013, pp. 667-675. https://aclweb.org/anthology/R/R13/R13-1087.pdf
8. Nikolova, I., I. Temnikova and G. Angelova. Enriching Patent Search with External Keywords: a Feasibility Study. In: Angelova, G., K, Bontcheva, and R. Mitkov, Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2013), September 2013, Hissar, Bulgaria, published by Incoma Ltd., Shoumen, Bulgaria, ISSN 1313-8502, 2013, pp. 525-531. https://aclweb.org/anthology/R/R13/R13-1069.pdf
9. Temnikova, I., W. Baumgartner Jr., N. Hailu, I. Nikolova, T. McEnery, A. Kilgarriff, G. Angelova, and K. B. Cohen. Sublanguage Corpus Analysis Toolkit: A tool for assessing the representativeness and sublanguage characteristics of corpora. In: Proceedings of LREC-2014, the 9th Int. Conf. on Language Resources and Evaluation, 26-31 May 2014, Reykjavik, Iceland, 2014, pp. 1714-1718, http://www.lrec-conf.org/proceedings/lrec2014/pdf/675_Paper.pdf
10. Nikolova, I., D. Tcharaktchiev, S. Boytcheva, Z. Angelov and G. Angelova. Applying Language Technologies on Healthcare Patient Records for Better Treatment of Bulgarian Diabetic Patients. In: G. Agre et al. (Eds.): AIMSA 2014, Springer Int. Publ. Switzerland, Lecture Notes in Artificial Intelligence 8722, 2014, pp. 92–103.

Следният постер описва резултатите от състезанието i2b2 и беше представен от съавтора Kevin Bretonnel Cohen на семинара The Sixth i2b2 Shared Task and Workshop Challenges in Natural Language Processing for Clinical Data: Temporal Relations, 2 November, 2012, Chicago, USA:

Nikolova, I., S. Boytcheva, G. Angelova and K. B. Cohen. Temporal expressions in clinical text: Event recognition and time expressions.
http://lml.bas.bg/~iva/docs/i2b2_2012_poster_Nikolova.pdf

Списък на цитиранията на публикации по дисертацията

- Nikolova, I. and G. Angelova. Identifying Relations between Medical Concepts by Parsing UMLS Definitions. In: Andrews, S., S. Polovina, R. Hill and B. Akhgar (Eds.), Conceptual Structures for Discovering Knowledge, Proc. of ICCS-2011, the 19th Int. Conference on Conceptual Structures, UK, July 2011, Springer, Lecture Notes in Artificial Intelligence 6828, 2011, pp. 173-186, DOI: 10.1007/978-3-642-22688-5_13

  2 цитата:

  - Ayisha Noori V. K, P. C. Reghu Raj. Extraction of Disease Relationship from Medical Records: Vector Based Approach. In: International Journal of Latest Trends in Engineering and Technology (IJLTET), Vol. 3 Issue2 November 2013, pp. 24-33, `http://ijltet.org/wp-content/uploads/2013/11/4.pdf`

  - Muhammad Zubair Asghar, Maria Qasim, Bashir Ahmad, Shakeel Ahmad, Aurangzeb Khan, Imran Ali Khan. Health miner: opinion extraction from user generated health reviews. International Journal of Academic Research Part a; 2013; 5(6), 279-284.

- Nikolova, I., G. Angelova, D. Tcharaktchiev and S. Boytcheva. Medical Archetypes and Information Extraction Templates in Automatic Processing of Clinical Narratives, In Proceedings of ICCS 2013, 10-12 January, Mumbai, India, Conceptual Structures for STEM Research and Education, Springer, Lecture Notes in Computer Science Volume 7735, 2013, pp 106-120.

  1 цитат:

  - Shalini Bhartiya and Deepti Mehrotra. Exploring Interoperability Approaches and Challenges in Healthcare Data Exchange. In: Proc. Smart Health, Springer, Lecture Notes in Computer Science Volume 8040, 2013, pp 52-65.

- Temnikova, I., W. Baumgartner Jr., N. Hailu, I. Nikolova, T. McEnery, A. Kilgarriff, G. Angelova, and K. B. Cohen. Sublanguage Corpus Analysis Toolkit: A tool for assessing the representativeness and sublanguage characteristics of corpora. In: Proceedings of LREC-2014, the 9th Int. Conf. on Language Resources and Evaluation, 26-31 May 2014, Reykjavik, Iceland, 2014, pp. 1714–1718.

  1 цитат:

  - Reinhard Rapp, Using Collections of Human Language Intuitions to Measure Corpus Representativeness, Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: pages 2117–2128, Dublin, Ireland, August 23-29 2014

Апробация на резултатите

Резултати от дисертацията са докладвани на следните конференции:

International Conference on Conceptual Structures (ICCS 2011), Derby, UK
6th Annual SEE Doctoral Student Conf. (SEERC 2011), Thessaloniki, Greece
European Chap. of the Ass. for Comp. Linguistics (EACL 2012), Avignon, France
Student Research Workshop ass. with the Conf. of the North American Chapter of the Ass. for Comp. Linguistics HLT (NAACL 2012), Montreal, Canada
Int. Conference on Conceptual Structures (ICCS 2013), Mumbai, India
Recent Advances in Natural Language Processing (RANLP 2013), Hissar, Bulgaria
Artificial Intelligence: Methodology, Systems and Applications (AIMSA 2014), Varna, Bulgaria - награда за "Най-добра статия на конференцията".

Участие в национални и международни проекти

Проект DO 02-292/Декември 2008 "Ефективно търсене на концептуални шаблони с приложения в медицинската информатика (ЕВТИМА)", финансиран от Националния фонд за научни изследвания през 2009-2012.

Проект BG051PO001-3.3.04/40: Изграждане на висококвалифицирани млади изследователи по съвременни информационни технологии за оптимизация, разпознаване на образи и подпомагане вземането на решения.

Проект No 216130 / 1.05.2010 - 31.07.2011: Сигурност на пациента чрез интелигентни процедури в лечението (PSIP+)

Проект BG051PO001/3.3-05 „НАУКА И БИЗНЕС" на МОМН - лична стипендия за едномесечно обучение в Денвър, Колорадо - University of Colorado School of Medicine. Работа върху извличане на събития от медицински епикризи на пациенти, диагностицирани с различни заболявания.

Проект No. 316087 "Съвременните пресмятания в полза на иновацията (AComIn)", финансиран от Европейската Комисия в FP7 Capacity Programme през 2012–2016.

Декларация за оригиналност на резултатите

Декларирам, че настоящата дисертация съдържа оригинални резултати, получени при проведени от мен научни изследвания {с подкрепата и съдействието на научния ми ръководител}. Резултатите, които са получени, описани и/или публикувани от други учени, са надлежно и подробно цитирани в библиографията.

Настоящата дисертация не е прилагана за придобиване на научна степен в друго висше училище, университет или научен институт.

Подпис:

Списък на използвани съкращения и означения

Термини

анализатор на части на речта Софтуерна програма/модул за обработка на естествен език, която определя частите на речта за думите в текста.

база за сравнение Характеристиките на базовата система, която се надгражда и служи за сравнение, за да се оцени подобрението в текущата система. От английската дума baseline.

чънкер Софтуерна програма/модул за обработка на естествен език, която прави плитък синтактичен анализ на текста. От английски - chunker.

дependентно дърво Дърво, представящо синтактичния анализ на изречение,чрез зависимости между думи. При дependентните връзки глаголът се счита за структурен център на цялата структура на изречението. Всички други синтактични елементи (напр. думи) са или директно,или индиректно зависими от глагола. От английски - dependency tree.

домейн Предметна област. Домейна/предметната област на изследването се определят от жанра и спецификата на документите, които се обработват - новинарски статии в различни категории; медицински текстове - епикризи, научни статии, листовки на лекарства и др. От английски - domain.

конкордансер Софтуерна програма/модул за обработка на естествен език, която търси съвпадения на търсена дума в текста и извежда като резултат обкръжението на думата в прозорец няколко думи/токъни (обикновено 5-7) наляво и надясно. От английски - concordancer.

конституентно дърво Дърво, представящо синтактичния анализ на изречение чрез словосъчетания. От английски - constituent tree.

конституент Под конституент разбираме словосъчетание. От английски - constituent.

корен Под корен разбираме част от строежа на думата. Корен е онази част от думата, която изразява основното ѝ лексикално значение.

лексикално затваряне Лексикално затваряне по отношение на даден признак при анализ на текст се наблюдава, когато при увеличаване на количеството на текста, спират да се появяват нови стойности на характеристиките на текста. Например не се наблюдават нови части на речта за думите; не се наблюдават нови последователности от морфологични маркери и др. От английски - lexical closure.

лематизатор Софтуерна програма/модул за обработка на естествен език, която определя лемата на думата. От английски - lemmatizer.

лема Основна форма на думата.

концептуален модел на предметна област Описва предметната област чрез основни понятия, които я характеризират и връзките между тези понятия.

онтология В компютърните науки онтология се нарича представянето на света в системата. Съществуват и онтологии с общо предназначение (като Cyc), но широко практикуван подход е света да

се моделира според целите на конкретно (компютърно) приложение.

парсер Софтуерна програма/модул за обработка на естествен език, която прави дълбок синтактичен анализ на текста. От английски - parser.

покриваемост Една от мерките, с която се измерва успеваемостта на алгоритмите/системите при задачите за извличане на информация, е покриваемостта. Тя се изчислява като отношение на правилно изведените единици спрямо всички, които са могли да бъдат изведени като отговарящи на критерия на търсене. От английски - recall.

предметна област Вж. домейн.

речник Под речници в тази дисертация разбираме списъци от термини, понятия, думи със специфична роля в дадената задача като например стоп-думи и т.н.

сигнални думи Това са думи, които се срещат в контекста на израза, който се опитваме да разпознаем. Сигналните думи може и да означават както наличието на дадени изрази, така и игнориране

стемер Софтуерна програма/модул за обработка на естествен език, която определя корена на думата. От английски - stemmer.

стоп-думи Най-често срещаните думи в текста, които не носят собствен смисъл и най-често са функционални (предлози,съюзи и т.н.). При някои подходи в обработка на естестествен език тези думи се игнорират и се работи само над думите със собствено значение. От английски - stop words.

точност Една от мерките, с която се измерва успеваемостта на алгоритмите/системите при задачите за извличане на информация е точността. Тя се изчислява като отношение на извлечените

вярно единици спрямо всички извлечени единици. От английски - precision.

токънайзер Софтуерна програма/модул за обработка на естествен език, която определя границите на буквено-цифровите низове в текста (думи и други низове). От английски - tokenizer.

токънизация Процес на изпълнение на програма токънайзер с цел определяне границите на буквено-цифровите низове в текста. От английски - tokenization.

токън Символен низ, който не съдържа разделители. Такива низове са думите, числата, буквено-цифрови низове и други. Разделителите на токъните могат да са различни в зависимост от задачата. Напр. интервал, определен препинателен знак, всички препинателни знаци без малко тире и др. От английски - token.

торба от думи Това е представяне на текста в матричен вид, при което редовете съответстват на думите, които се срещат в колекцията, стълбовете - на документите в колекцията, а във всяка клетка се записва броят срещания на съответната дума в съответния документ. От английски - bag of words.

F-measure Една от мерките, с която се измерва успеваемостта на алгоритмите/системите при задачите за извличане на информация е F-measure. Тя се хармонично средно на точността и покриваемостта.

Означения

ИИ Изкуствен интелект.

МС Машинно самообучение.

МЕИД Междуезиково извличане на документи.

МКБ / ICD Международна класификация на болестите (International Classification of Diseases). http://www.who.int/classifications/icd/en/

ОЕЕ Обработка на естествен език.

ПЗ Пациентски записи.

СС Семантични системи.

ССТ Семантични системи, обработващи текст.

CRF Conditional Random Fields

CUI Concept Unique Identifier - уникален идентификатор на понятията в UMLS.

MEDLINE® Основната онлайн база от библиографични цитати на Националната библиотека за медицина на САЩ, която осигурява международен достъп до списанията на био медицинска тематика по света.

MeSH Medical Subject Headings - Названия на медицински понятия. http://www.nlm.nih.gov/mesh/

NLM National Library of Medicine - Национална библиотека за медицина на САЩ. http://www.nlm.nih.gov/

PubMED® PubMed съдържа повече от 24 милиона цитати на био-медицинска литература от MEDLINE.
http://www.ncbi.nlm.nih.gov/pubmed

SNOMED CT® SNOMED Clinical Terms® - най-изчерпателна многоезична терминологична банка на клиничните термини в здравеопазването. Съдържа основната терминология, описваща електронно здравно досие и съдържа повече от 311 000 активни понятия с уникално значени и формални дефиниции, базирани на логически форми, организирани в йерархии.
http://www.ihtsdo.org/snomed-ct/

UMLS® Unified Medical Language System® - многоезиков ресурс, който интегрира ключова терминология, класификации, стандарти за кодиране и асоциирани ресурси с цел да се създадат по-ефективни и съвместими био-медицински информационни системи и услуги, включително електронни здравни записи. http://www.nlm.nih.gov/research/umls/

UTS UMLS Terminology Services - сървъри, предоставящи възможност за отдалечен достъп до базата на UMLS. https://uts.nlm.nih.gov/home.html

WordNet® WordNet е голяма лексикална база на английски език. Съществителни, глаголи, прилагателни и наречия са групирани в множества от когнитивни синоними, всяко от които изразява отделно понятие. http://wordnet.princeton.edu/

Списък на фигурите

1.1 Структура на семантична система, обработваща текст. 12

2.1 Примери, демонстриращи семантични релации. 25
2.2 Пример, демонстриращ обобщение. 26
2.3 Извадка от модел на предметна област. 26
2.4 Извличане на парафрази. 31
2.5 Извличане на термини от колекция с документи, базирано на честота на срещане на думите в текста. 32
2.6 Примерна епикриза. 34
2.7 Процес на създаване на модела на предметната област. 45
2.8 RelEx резултати, илюстриращи депендентните връзки, които подпомагат извличането на релации. 52
2.9 Депендентен граф, обясняващ извличането на релация **IS-A** по правило 2.3.1 53
2.10 Интеграция на характеристики и извличане на нови релации. . . . 55

3.1 Общ шаблон на отношенията, представляващи описание на симптом. 66
3.2 Шаблон на релацията *ниво на кръвната захар*. 66
3.3 Добавяне на нови връзки между корените. 68
3.4 Извличане на речници на събития, темпорални изрази, съответни семантични типове и други от обогатените с лингвистична и семантична информация епикризи. 80
3.5 Обща схема на извличането на събития от епикризи. 81

Списък на таблиците

2.1 Множество от термини, които се срещат с feet в ПЗ. 35
2.2 Вътрешно представяне на характеристиките на термините като наредени четворки. 37
2.3 Резултати от прилагането на правила 2.2.1–2.2.8 върху тестовите данни. 40
2.4 Точност и покриваемост на правила 1–8. 41
2.5 Извлечени дефиниции от UMLS за термина ”pulse”. 48
2.6 Нормализация на дефинициите, извлечени от UMLS 49
2.7 Коригиране на безглаголни дефиниции, извлечени от UMLS . . . 50
2.8 Грешно разрешаване на многозначността на ниво част на речта . 57

3.1 BIO анотация за полидипсо-полиуричен синдром. 64
3.2 BIO анотация за промяна в теглото. 65
3.3 Фаза 1 - правила за последователности от думи след определяне на корените. 68
3.4 Разпознаване на резултати от лабораторни тестове. 70
3.5 Лява и дясна граница на описанията. 71
3.6 3. Оценяване на фаза 1. Резултати от подходите базирани на правила и на МС. 72
3.7 4. Резултати от отделните фази на прилагане на правилата 73
3.8 Описателна статистика за учебните и тестови данни 79
3.9 Маркиране/разпознаване на събития със *Система 2* 85
3.10 Маркиране/разпознаване на събития със *Система IV*. 88
3.11 Оценяване на извличането на събития върху тестовото множество от данни с програми предоставени от i2b2. 89

4.1 Резултати от Експеримент 1 и 2 с ръчно подбрани характеристики. 97
4.2 Резултати от експерименти 3a и 3b - автоматичен подбор на характеристики, ползване на биграми и триграми. 97
4.3 Експерименти 4a и 4b с MaxEnt и всички текстови характеристики, биграми и триграми 98

Библиография

[1] Sophia Ananiadou and J. Mcnaught. *Text Mining for Biology And Biomedicine*. Artech House, Inc., 2005.

[2] Alan R. Aronson. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of AMIA Symposium*, pages 17–21, 2001.

[3] Alan R. Aronson and François-Michel Lang. An overview of metamap: historical perspective and recent advances. *J Am Med Inform Assoc.*, 17(3):229–236, May-Jun 2010.

[4] Svetla Boytcheva, Ivelina Nikolova, Elena Paskaleva, Galia Angelova, Dimitar Tcharaktchiev, and Nadya Dimitrova. Obtaining Status Descriptions via Automatic Analysis of Hospital Patient Records. *Informatica*, 34(3):269–278, 2010.

[5] Svetla Boytcheva, Albena Strupchanska, Elena Paskaleva, and Dimitar Tcharaktchiev. Some Aspects of Negation Processing in Electronic Health Records. In *International Workshop LSI in the Balkan Countries*, 2005.

[6] Philip Bramsen, Pawan Deshpande, Yoong Keok Lee, and Regina Barzilay. Finding temporal order in discharge summaries.

[7] Paul Buitelaar, Daniel Olejnik, and Michael Sintek. A protégé plug-in for ontology extraction from text based on linguistic analysis. In Christoph. Bussler, John Davies, Dieter Fensel, and Rudi Studer, editors, *The Semantic Web: Research and Applications*, volume 3053 of *Lecture Notes in Computer Science*, pages 31–44. Springer Berlin Heidelberg, 2004.

[8] Wendy Chapman, Will Bridewell, Paul Hanbury, Gregory Cooper, and Bruce Buchanan. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. *Journal of Biomedical Informatics*, 34:301–310, 2001.

[9] Wendy Chapman and Kevin Bretonnel Cohen. Current issues in biomedical text mining and natural language processing. *Journal of Biomedical Informatics*, 42(5):757–759, October 2009.

[10] Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL 2002)*, 2002.

[11] Hamish Cunningham, Diana Maynard, Kalina Bontcheva, Valentin Tablan, Niraj Aswani, Ian Roberts, Genevieve Gorrell, Adam Funk, Angus Roberts, Denica Damljanovic, Thomas Heitz, Mark A. Greenwood, Horacio Saggion, Johann Petrak, Yaoyong Li, and Wim Peters. *Text Processing with GATE (Version 6)*. 2011.

[12] Berry de Bruijn, C Cherry, S Kiritchenko, Joel Martin, and X Zhu. NRC at i2b2: one challenge, three practical tasks, nine statistical systems, hundreds of clinical records, millions of useful features. In *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data*, 2010.

[13] Kerstin Denecke. Enchancing knowledge representations by ontological relations. In K. Andersen et al., editor, *eHealth Beyond the Horizon - Get IT There, Proceedings of MIE 2008, Goteborg 2008*, volume 136, pages 791–796. IOS Press, 2006.

[14] Bill W. Dolan, Chris Quirk, and Chris Brockett. Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources. In *COLING '04 Proceedings of the 20th international conference on Computational Linguistics*, 2004.

[15] Rezarta Islamaj Doğan and Zhiyong Lu. An inference method for disease name normalization. In *Proceedings of the AAAI 2012 Fall Symposium on*

Information Retrieval and Knowledge Discovery in Biomedical Text, pages 8–13, 2012.

[16] Christiane Fellbaum, editor. *WordNet: An electronic lexical database*. MIT Press, Cambridge, MA, 1998.

[17] Carol Friedman, Philip O. Alderson, John H. M. Austin, James J. Cimino, and Stephen B. Johnson. A General Natural-Language Text Processor for Clinical Radiology. *JAMIA*, Mar-Apr, 1(2):161–74, 1994.

[18] Katrin Fundel, Robert Küffner, and Ralf Zimmer. Relex—relation extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371, January 2007.

[19] Dennis Grinberg, John Lafferty, and Daniel Sleator. A robust parsing algorithm for link grammars. Technical Report CMU-CS-95-125, Carnegie Mellon University, 1995.

[20] H Gurulingappa, M Hofmann-Apitius, and J Fluck. Concept identification and assertion classification in patient health records. In *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data*, 2010.

[21] Udo Hahn, Martin Romacker, and Stefan Schulz. MEDSYNDIKATE - a Natural Language System for the Extraction of Medical Information from Findings Reports. *Int J Med Inf*, 67(1-3):63–74, 2002.

[22] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November 2009.

[23] Henk Harkema, John Dowling, Tyler Thornblade, and Wendy Chapman. ConText: An Algorithm for Determining Negation, Experiencer, and Temporal Status from Clinical Reports. *J Biomed Inf*, 42(5):839–51, 2009.

[24] Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational Linguistics - Volume 2*, COLING '92, pages 539–545, Stroudsburg, PA, USA, 1992. Association for Computational Linguistics.

[25] Min Jiang, Yukun Chen, Mei Liu, Trent Rosenbloom, Subramani Mani, Joshua C Denny, and Hua Xu. Hybrid approaches to concept extraction and assertion classification - Vanderbilt's systems for 2010 I2B2 NLP Challenge. In *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data*, 2010.

[26] Antonio Jimeno, Sylvain Gaudan Rafael Berlanga Ernesto Jimenez-Ruiz, Vivian Lee, and Dietrich Rebholz-Schuhmann. Assessment of disease named entity recognition on a corpus of annotated sentences. *BMC Bioinformatics*, 9(Suppl 3), 2008.

[27] Ning Kang, Rogier Barendse, Zubair Afzal, Bharat Singh, Martin Schuemie, Erik van Mulligen, and Jan Kors. Erasmus MC approaches to the i2b2 Challenge. In *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data*, 2010.

[28] Ning Kang, Bharat Singh, Zubair Afzal, Erik M. van Mulligen, and Jan A. Kors. Using rule-based natural language processing to improve disease normalization in biomedical text. *J Am Med Inform Assoc.*, pages 876–881, October 2012.

[29] Helena Karsten and Hanna Suominen. Mining of clinical and biomedical text and data: Editorial of the special issue. *International Journal of Medical Informatics*, 78(12):786–787, 2009.

[30] Zornitsa Kozareva and Andrés Montoyo. Paraphrase identification on the basis of supervised machine learning techniques. In Tapio Salakoski, Filip Ginter, Sampo Pyysalo, and Tapio Pahikkala, editors, *Advances in Natural Language Processing: 5th International Conference on NLP (FinTAL 2006)*, volume 4139 of *Lecture Notes in Computer Science*, pages 524–533. Springer Berlin Heidelberg, 2006.

[31] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[32] John D. Lafferty, Danny Sleator, and Davy Temperley. Grammatical trigrams: a probabilistic model of link grammar. In *Proceedings of the AAAI Fall Symposium on Probabilistic Approaches to Natural language*, 1992.

[33] Robert Leaman, Islamaj Rezarta Doğan, and Zhiyong Lu. Dnorm: disease name normalization with pairwise learning to rank. *Bioinformatics*, 29(22), October 2013.

[34] Andrew Kachites McCallum. Mallet: A machine learning for language toolkit.

[35] Tony McEnery and Andrew Wilson. *Corpus Linguistics*. Edinburgh University Press, 2nd edition, 2001.

[36] Rada Mihalcea, Courtney Corley, and Carlo Strapparava. Corpus-based and knowledge-based measures of text semantic similarity. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*, AAAI'06, pages 775–780. AAAI Press, 2006.

[37] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, November 1995.

[38] Preslav Nakov. BulStem: Design and Evaluation of Inflectional Stemmer for Bulgarian. In *Proc. of Workshop on Balkan Language Resources and Tools (1st Balkan Conference in Informatics)*, 2003.

[39] Ivelina Nikolova. Unified extraction of health condition descriptions. In *Proceedings of the NAACL HLT 2012 Student Research Workshop*, pages 23–28, Montréal, Canada, June 2012. Association for Computational Linguistics.

[40] Sergei Nirenburg, Marjorie McShane, Margalit Zabludowski, Stephen Beale, and Craig Pfeifer. Ontological semantic text processing in the biomedical domain. Working Paper 03-05, University of Maryland Baltimore County, Institute for Language and Information Technologies, 2005.

[41] Martin F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[42] Long Qiu, Min-Yen Kan, and Tat-Seng Chua. Paraphrase recognition via dissimilarity significance classification. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 18–26, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[43] Hans Reichenbach. *Elements of symbolic logic*. The Macmillan Company, 1947.

[44] Bryan Rink, Sanda Harabagiu, and Kirk Roberts. Extraction of medical concepts, assertions, and relations from discharge summaries for the fourth i2b2/VA shared task. In *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data*, 2010.

[45] Ferdinand de Saussure. *Course in general linguistics*. New York : Philosophical Library, 1959.

[46] Guergana Savova, Philip Ogren, Patrick Duffy, James Buntrock, and Christopher Chute. Mayo Clinic NLP System for Patient Smoking Status Identification. *JAMIA*, 15(1):25–28, 2008.

[47] Guergana K. Savova, James J. Masanz, Philip V. Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C. Kipper-Schuler, and Christopher G. Chute1. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *JAMIA*, 17(5):507в„––513, 2010.

[48] Alexander Schutz and Paul Buitelaar. Relext: A tool for relation extraction from text in ontology extension. In Yolanda Gil, Enrico Motta, V.Richard Benjamins, and MarkA. Musen, editors, *The Semantic Web – ISWC 2005*, volume 3729 of *Lecture Notes in Computer Science*, pages 593–606. Springer Berlin Heidelberg, 2005.

[49] Matthew S. Simpson and Dina Demner-fushman. Biomedical text mining: a survey of recent progress, 2012.

[50] Daniel D. Sleator and Davy Temperley. Parsing English with a link grammar. In *Proc. Third International Workshop on Parsing Technologies*, pages 277–292, 1993.

[51] Karen Sparck-Jones and Peter Willett, editors. *Readings in Information Retrieval*. Morgan Kaufmann, 1997.

[52] Michael Q. Stearns, Colin Price, Kent A. Spackman, and Amy Y. Wang. Snomed clinical terms: Overview of the development process and project status. In *Proceedings of the AMIA Symposium*, page 662вЂ“666, 2001.

[53] Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *J Am Med Inform Assoc.*, 20(5):806–813, Sep-Oct 2013.

[54] Irina Temnikova and Kevin Cohen. Recognizing sublanguages in scientific journal articles through closure properties. In *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing*, pages 72–79, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[55] Irina Temnikova, Ivelina Nikolova, William A. Baumgartner, Galia Angelova, and K. Bretonnel Cohen. Closure properties of bulgarian clinical text. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 667–675, Hissar, Bulgaria, September 2013. INCOMA Ltd. Shoumen, BULGARIA.

[56] Ozlem Uzuner, Brett South, Shuying Shen, and Scott DuVall. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *JAMIA*, pages i116–i124, 2011.

[57] Cornelis J. van Rijsbergen, Stephen E. Robertson, and Martin F. Porter. New models in probabilistic information retrieval. 1980.

[58] Mark Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Jessica Moszkowicz, and James Pustejovsky. Semeval 2007 task 15: Tempeval temporal relation identification. 2007.

[59] Mark Verhagen, R Sauri, T Caselli, and James Pustejovsky. Semeval-2010 task 13: Tempeval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 57–62, 2010.

[60] Špela Vintar, Paul Buitelaar, and Martin Volk. Semantic relations in concept-based cross-language medical information retrieval. In *ECML/PKDD Workshop on Adaptive Text Extraction and Mining (ATEM)*, 2003.

[61] Špela Vintar, Ljupčo Todorovski, Daniel Sonntag, and Paul Buitelaar. P.: Evaluating context features for medical relation mining. In *In: ECML/PKDD Workshop on Data*, 2003.

[62] S. Visa, A. Ralescu, and M. Ionescu. Investigating Learning Methods for Binary Data. In *Proc. Fuzzy Information Processing Society, 2007. NAFIPS '07. Annual Meeting of the North American*, pages 441–445, 2007.

[63] Sthephen Wan, Mark Dras, Robert Dale, and Cécile Paris. Using Dependency-based Features to Take the "Para-farce"out of Paraphrase. In *Australasian Language Technology Workshop 2006 (ALTW 2006)*, pages 131–138, 2006.

[64] Sander Wubben, Antal van den Bosch, Emiel Krahmer, and Erwin Marsi. Clustering and matching headlines for automatic paraphrase acquisition. In *Proceedings of the 12th European Workshop on Natural Language Generation*, ENLG'09, pages 122–125, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[65] Yitao Zhang and Jon Patrick. Paraphrase identification by text canonicalization. In *Proceedings of the Australasian Language Technology Workshop 2005*, pages 160–166, 2005.

[66] Pierre Zweigenbaum and Dina Demner-Fushman. Advanced literature-mining tools. In D. Edwards, J. Stajich, and D. Hansen, editors, *Bioinformatics: Tools and Applications*, pages 347–380. Springer, 2009.

[67] Pierre Zweigenbaum, Dina Demner-Fushman, Hong Yu, and Kevin B. Cohen. Frontiers of biomedical text mining: Current progress. *Briefings in Bioinformatics*, 8(5):358–375, 2007.