БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ
ИНСТИТУТ ПО ИНФОРМАЦИОННИ И КОМУНИКАЦИОННИ ТЕХНОЛОГИИ

Ивелина Мирчева Николова

ПРИЛОЖЕНИЕ НА ОБРАБОТКАТА НА ЕСТЕСТВЕН ЕЗИК ЗА ИЗГРАЖДАНЕ НА СЕМАНТИЧНИ СИСТЕМИ

А В Т О Р Е Ф Е Р А Т

на дисертация 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.

Дисертационният труд е обсъден и допуснат до защита на разширено заседание на секция „Лингвистично моделиране“ на ИИКТ-БАН, състояло се на 25 септември 2014 г.

Дисертационният труд съдържа 129 страници, в които 15 фигури, 22 таблици и 8 страници литература, включваща 67 заглавия.

Защитата на дисертацията ще се състои на ..... г. от 14:00 часа в зала ..... на Института по ..... - БАН на открито заседание на научно жури в състав:

1. ...
2. ...
3. ...
4. ...
5. ...

Материалите за защитата са на разположение на интересуващите се в стая 215 на ИИКТ-БАН, ул. „Акад. Г. Бончев“, блок 25А, град София.

Автор: Ивелина Мирчева Николова

Заглавие: Приложение на обработката на естествен език за изграждането на семантични системи

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, 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.

Обща характеристика на дисертационния труд

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.

Методология на изследването

Настоящите изследвания се базират на резултати от фундаментални научни изследвания в областта на обработката на естествения език и семантичните системи. Комбинират се подходи базирани на правила, както и на статистически анализ и машинно самообучение. При анализ на текста се ползват различни негови характеристики - повърхнинни, базирани на речници, контекстни, лексикални, морфо-синтактични и семантични.

Настоящата дисертация се състои от увод, 3 глави, заключение и списък на цитираната литература. Основното съдържание е поместено на 111 страници, придружено с 15 фигури и 22 таблици. Списъкът на цитираната литература включва 67 заглавия.

Глава 1. Обзор на основните резултати в областта

Автоматичната обработка на естествения език е бурно-развиваща се област, тъй като днес се приема за почти задължително, че компютрите трябва да бъдат интелигентни, да разбират текст и реч, да превеждат, да търсят вместо нас в Интернет и да отговарят на въпроси и т.н. Настоящата дисертация е посветена на автоматичния анализ на медицински текстове и интеграция на разработените алгоритми и прототипи в семантични системи за обработка на големи обеми от пациентски данни.

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. Модулите, които обичайно съставят една ССТ са 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. 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. 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 също е във фокуса на тази дисертация. Тук първо се прави общ анализ на текста за автоматич-

но определяне на лексикалните и морфосинтактичните му характеристики. 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 повърхнинни, морфосинтактични признаци , структурни, контекстни, семантични 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 са описани по-подробно в следващите глави.

Автоматично разпознаване на парафрази

Определянето на близост между документи и изрази, и в частност разпознаването на парафрази е в основата на много задачи при съвременната обработка на естествен език (ОЕЕ) и семантичните технологии. От успешното разрешаване на този проблем зависи качеството на анализа и като следствие – извличането и структурирането на информация от свободен текст. Парафразите могат да бъдат изрази с различна дължина. Това са както синоними на термини, означаващи едно понятие, изразени с една или няколко думи, до изречения, съдържащи същата информация или заглавия на статии на една и съща тема.

Днес при разпознаването на парафрази се използват както подходи, базирани на правила, така и машинно самообучение и хибридни подходи. В [24] 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 [27], идентифицират понятията и прилагат мерки за лексикална близост върху конституентите, които ги съдържат. Друг класически подход е изчисляването на близост между парафрази чрез мерки за близост на вектори. В [39] се ползват вариации на $tf - idf$ за представяне на документи, а в [40] 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.

В [22] са публикувани едни от най-добрите резултати за точност, която е постигната досега при разпознаване на парафрази на заглавия. Чрез анализ на лексикални и семантични характеристики на текста се постига точност от 76.6% и покриваемост - 79.6%. Този подход има малко по-ниска покриваемост в сравнение с други подобни, например в [38] методите работят с 83% покриваемост, но с малко по-ниска точност - 75.6%. Споменатите подходи разпознават парафрази чрез мерки за лексикална близост, семантично представяне и различни техники за машинно самообучение, които ние също използваме.

Типична задача, включваща разпознаване на парафрази и синоними, е откриването и нормализирането на имена на заболявания, симптоми и анатомични органи. Сложността произтича от няколко фактора, които са характерни за текстове в областта на биомедицината. Имената на заболяванията често съдържат думи от гръцки и латински, например "lymphogranulomatosis". Те могат да бъдат композирани от категория на болестта и модификатор, например "рак на гърдата", или симптом "cat-eye syndrom", вид лечение "инсулинозависим захарен диабет (тип I)", агент, причинител на заболяването "стрептококова инфекция", етимология идваща от биомолекулярната специфика на заболяването "идеопатична парапротеинемия", епонимия "синдром на Сьогрен" и др. Освен това модификатори се използват и за определяне на състоянието на заболяването, а не само като част от името, какъвто е случаят с "обострен синузит".

Друго затруднение за откриване на парафразите са съкращенията, инфлексията на словоформите, ортографския запис, подредбата на думите в съставните имена и синонимите. 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 ([20], [9], [26]). Когато "нормализираме" съставни фрази, обик-

новено са полезни премахването на специфична ортография, свеждане на словоформите към основа и други техники за промяна на повърхностните "характеристики", но вариантите на изписване на имената от лексикона винаги поставят някакви ограничения. За тяхното преодоляване с машинно самообучение има сравнително малко изследвания, като едно от тях е [28].

Извличане на понятия и релации между тях

С развитието на областта "Представяне на знанията" в Изкуствения интелект става ясно, че при създаването на декларативни концептуални описания е неизбежно съприкосновението с естествения език. Класически определения за релации между понятията са предложени от Сосюр в [30]: синтагматични и асоциативни (парадигматични). Първите са валидни само в рамките на даден контекст, а асоциативните се базират на натрупан опит. В Изкуствения интелект, 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 етикети от текст, а ребрата са връзките между тези понятия. Специален пример за такава мрежа е WordNet [25], [10]. Възникват логическите формализми за представяне на знания. Създават се големи, ръчно-направени онтологии: Cyc, OpenMind Common Sense, MindPixel, FreeBase. Увеличаващите се ресурси в електронен формат и желанието за по-бърз напредък водят до първите опити за автоматично създаване на концептуални ресурси. В [18] се описва метод за автоматично извличане на лексикални релации на хипонимия от неструктуриран текст. Методът не предполага предварително кодиране на концептуални единици и е приложим за текстови колекции от разнообразни области. Авторите на [5] 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 описваща неврологията, 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 текстове в предметната област. RelExt извлича глаголи и техните граматични аргументи (напр. термини) и изчислява съответните релации чрез комбиниране на лингвистични и статистическа обработка [32].

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

Система с подобно име – RelEx – подпомага извличането на релации от свободен текст [12]. RelEx прави синтактичен анализ и извежда дependентни дървета на входните изречения, като прилага набор от правила върху тях с цел идентифициране на релации. През 2007 RelEx е оценена върху 1 млн резюмета от Medline, отнасящи се за релации между гени и протеини, и е извлякла около 150 000 релации с точност 80% и 80% покриваемост [12].

В [8] се показва начин за използване на онтологичните релации от UMLS[2] (най-големият ресурс от медицински номенклатури, събирани от Националната библиотека по медицина на САЩ от 1986 г.), за да се произведе автоматично структурирано представяне на неструктурирани медицински записи. Използват се връзките между понятията и мрежата от семантични типове на UMLS, за да се разпознаят допълнителни семантични релации, които да подпомогнат процеса на структуриране. Има много изследвания за извличане на релации, но ние се фокусираме върху извличане от кратки медицински дефиниции и описване на ограничения (правила за филтриране), които помагат да се уточнят и обяснят откритите релации. Статията [35] представя метод за филтриране на релации и метод за откриване на нови такива, идентифицирани при междуезиково извличане на документи (МЕИД), като използва семантични релации в медицината. База за сравнение са семантичните релации между медицински понятия в UMLS. Оценката е извършена над корпус от медицински резюмета на английски и немски език. 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 [36]. Авторите представят изследване за идентифициране и оценяване на контекстни характеристики и методи за машинно самообучение, за да се разпознаят медицински семантични релации 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.

В резюме, задачата за извличане на релации от биомедицински текстове се атакува чрез разнообразни техники, често в интегрирани подходи. Експериментите демонстрират значимостта на статистически извлечените правила за текстообработка, но също така използват лингвистичните характеристики и експертното знание; тази комбинация е съществена за разбирането на био-медицински език [6]. Средата OntoLT, която се ползва за извличане на онтологии от текст, предлага език за дефиниране на пред-

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

варителни условия (правила за съпоставяне) 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.

Автоматично структуриране на описания от пациентски записи

Задачата за структуриране на описания на електронни здравни записи е от съществено медицинско 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 OEE и по-точно извличане на информация (Information Extraction) в областта на биомедицината.

С увеличаване на обема на здравните записи в електронен формат естествено се налага те да бъдат ефективно съхранявани и обработвани с цел извличане на структурирани описания в тях. Става възможно те се ползват като: (*i*) източник за наблюдения, които могат да се правят само над голям обем данни; (*ii*) справочна информация за минали случаи, взаимодействия на лекарства, последователности от лечения и др.; (*iii*) изходен ресурс за създаване на регистри за конкретен тип заболявания. Тенденцията е вече разпространена в повечето държави от развития свят, но започва от САЩ, където се влагат най-много средства за разработването на такива системи и съответно най-успешните прототипи за структуриране на здравното състояние на пациентите работят предимно за записи на английски език.

Системата представена в [31] решава задачата за идентифициране на пациенти-пушачи чрез класификация на отделни изречения от статуса на пациента. Тя постига F-мярка 85,57%, като едно от ограниченията ѝ 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 винаги описани 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.

Статията [17] представя 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 отчети с F-мярка на разпознаване на състоянията както следва:

отрицание (75-95%), минали (22-84%), хипотетични (86-96%) и отнасящи се за пациента (100%) в зависимост от типа на документа. Нашият подход използва подобна идея – учим речници със "сигнализиращи" думи от данните и ги прилагаме за определяне на границите на изразите, които искаме да извлечем, но за разлика от [17] се фокусираме върху извличане на описания на симптоми, статуса им, стойности и евентуално отрицание.

Отрицанието е една от най-важните характеристики за разпознаване в медицинския текст. Наблюдения върху български клинични записи са представени в [4], където проблемът се третира чрез анализ на сигнализиращи изрази. Същият подход прилагаме и ние.

Много системи съдържат модули за извличане на отделни състояния или симптоми на пациента, но рядко се извлича завършен модел на всички състояния. Например MedLEE [11] и MEDSYNDIKATE [15] идентифицират статуса на състоянието и модифициращата информация като анатомично място, отрицание, промяна през времето. В [3] авторите извличат от български записи с висока точност статуса на кожата, крайниците, шията и щитовидната жлеза.

В тази дисертация от неструктуриран (свободен) текст се извличат описания на симптоми на пациента в ограничена предметна област, като се комбинират техники, базирани на правила с такива от машинното самообучение. Идентифицират се понятия и релации между тях.

Глава 2. Приложение на OEE за създаване на концептуални модели на предметна област

За да се направи задълбочен автоматичен анализ на текстови документи на определена тематика е необходим 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 дефинират основните понятия от областта и релациите между тях. Такива са онтологиите и базите от знание. Те подпомагат 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) на български език, създадени във връзка с болничен престой на диабетици. 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 диагностиката и лечението;

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

Освен понятията и 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 от типа **IS-A** и **AFFECTS**. Връзката **IS-A** или още известна като **дете-родител** свързва по-специфични с по-общи понятия, а връзката **засяга**, известна като **AFFECTS**, свързва заболявания или симптоми с органите, в които те причиняват патологични изменения.

Обогатявайки модела със семантични типове на понятията от областта, ние разширяваме възможностите за анализ на текста и за разбиране на фактите в него. 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.

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.

Тъй като не съществува единен концептуален или терминологичен ресурс на български език, описващ областта на диабета, 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 и фрагменти от концептуални описания, но това е трудоемка задача, поради което се налага да намерим полу-автоматични начини за извличането му от множество източници и работния корпус от документи.

2.1. Разпознаване на синоними и парафрази на понятията от наличната терминологична банка

Въведение

Разпознаването на парафрази е добре известен проблем, който се решава както в едноезиков, така и в многоезиков контекст. Лексикалното разнообразие, с което може да се изрази смисълът на дадено понятие, прави задачата сложна дори за хора. 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 (Unified Medical

Language System), които могат да са етикети на термини, използвани в нашия корпус от епикризи.

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 полу-автоматично да го обогатим с парафрази/синоними на термините в него. Така създаденият ресурс е основата на модела на предметната област. За разрешаване на този проблем демонстрираме подход за извличане на термини на английски от UMLS, които могат да са етикети на термини, използвани в нашия корпус от ПЗ.

Метод

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. Не винаги 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 (*iv*)).

Използваме подход **отдолу-нагоре** за построяване на концептуалния ресурс. Започваме с предварителна честотна обработка на корпуса от епикризи и с набор от статистически методи, техники за OEE и ръчно отсяване на значимите понятия извличаме списък на термините, употребени в корпуса. Създаваме *списък с термини, описващи крайниците* и *списък с термини-колокации от пациентски записи*, които впоследствие използваме в правила за извличане на парафрази.

За намиране на правилните съответствия (парафрази/ синоними) между термини от новосъздадения речник и термини на английски от UMLS прилагаме правила, създадени след анализ на учебния 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). Използваме частично от UMLS следните ресурси: *понятия*, *синонимни множества*, *семантични типове* и *релации*.

В резултат от моделиране на фразова структура на термините извличаме 8 правила за разпознаване на парафрази, които са базирани на лек-

сикална близост, такива, които са специфични за задачата, и правила базирани на контекста.

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. 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 отговора за една заявка). Термини, които не са получили нито едно предложение от UMLS са пренебрегнати. За тестване, ние отделихме 66 UMLS отговора и проверихме правилата върху тях. За всеки термин разглеждаме всички предложения от UMLS за него. Резултатите от филтрирането на предложенията от UMLS са представени на Таблица 1 .

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

Таблица 1. Резултати от прилагането на правилата върху тестовите данни.

Оценихме ръчно релевантността на етикетите от UMLS спрямо термините от тестовото множество. Оказа се, че от 66 UMLS етикета само 46 наистина съответстват на изходните 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 .

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

Таблица 2. Точност и покриваемост на правилата, разпознаващи парафрази на термини.

Резултатите са окуражаващи - те показват, че в затворена област с относително ясна терминология, подход, базиран на правила и метрики за лексикална близост, може да доведе до сравнително високи резултати. Правила 1, 5 и 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.

Резюме

Представихме подход за намиране на парафрази в междуезиков контекст. Извличаме парафрази на термини, описващи статуса на крайници в епикризи на диабетно-болни на български език. Предметната област е затворена и терминологията е доста добре дефинирана, това 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 знак, че можем да се опрем на съществуващи ресурси и не е нужно да ги създаваме 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 резултати. 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.

2.2. Разпознаване на релации между медицински понятия чрез обработка на дефиниции от UMLS

Въведение

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 като SNOMED CT[3] [33] и MeSH[4], интегрирани в UMLS, и могат да бъдат извлечени оттам, но те не са изчерпателни. Дори йерархичната релация **IS-A** не винаги е означена в тях.

Дефинирането на релации между понятия е сравнително труден процес, тъй като те отразяват връзки, които могат да са неявни и често зависещи от задачата. 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

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

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.

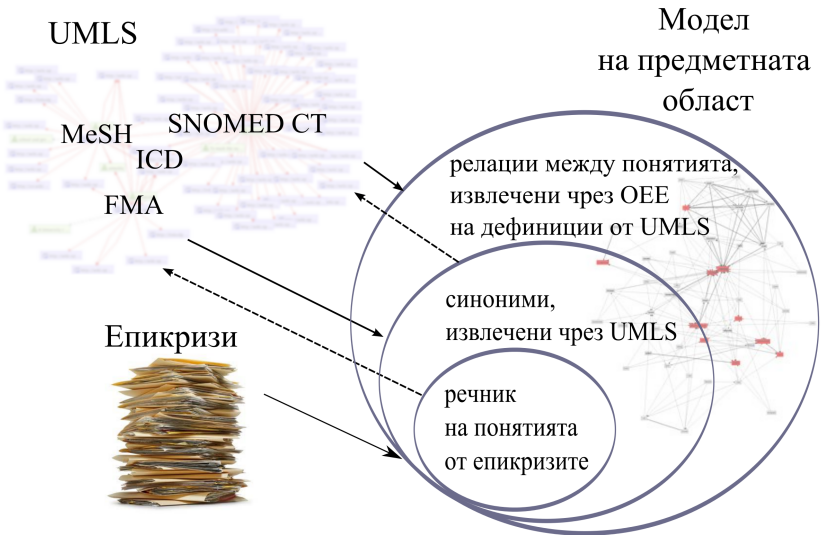
Дефиниция на задачата

Задачата, която си поставяме е да обогатим терминологичния речник, построен на предходния етап с връзки между понятията, както е показано на Фигура 1 . 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**. Изходният речник е на български, както и крайният резултат, а извличането на дефинициите и тяхната обработка се извършва на английски.

Метод

Текстовият корпус в нашия експеримент е колекция от дефиниции на термини, извлечени от UMLS Metathesaurus. Всяко понятие с уникален *CUI* (Concept Unique Identifier) представлява едно значение на даден термин и е дефинирано в някоя от стотиците интегрирани номенклатури на 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.

Прилагаме предварителна обработка на повърхностни характеристики на дефинициите, с което целим намаляване на възможностите за грешка в последващите фази. Ортографията и пунктуацията са от съществено значение за автоматичния синтактичен анализ и малки трансформации в тях могат да подобрят резултатите от анализа. Филтрирането на детайли, които нямат отношение към 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.



Фигура 1. Процес на създаване на модела на предметната област.

На избраните дефиниции правим синтактичен и семантичен анализ със системата RelEx. RelEx е синтактичен и семантичен анализатор. Той обработва текст на английски език и връща като резултат дependентните синтактични структури, които описват изреченията и семантичното обвързване между обектите, базирано на синтактични и семантични категории.

Релация IS-A

Извличаме нови релации от тип **IS-A**, като се базираме на извлечения дependентия граф на изреченията (извлечен от RelEx) и следвайки общоизвестни правила за композиция на смисъла. Вземаме предвид обичайната структура на английското изречение *подлог-сказуемо-допълнение* (ПСД) и факта, че дефинициите са кратки изречения.

Релация AFFECTS

Фокусираме се върху случаите, в които **AFFECTS** изразява връзка между усложнения/болести/симптоми и органа, който те засягат. Разглеждаме примерите, извлечени от дефинициите на усложненията на Diabetes Mellitus. Разпознаваме релацията по явно споменаване на основата affect* в обработваната дефиниция, т.е. трябва да са налични думите *affect*, *affects*, *affecting*, *affected* и т.н. За да извлечем релациите AFFECTS, представяме всяка дефиниция в лексико-семантична структура. Тази структура се построява на 5 стъпки.

Експерименти и резултати

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. Използвахме тези зависимости при изработване на правилата за извличане на релацията **IS-A**. 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, 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 една 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.

Заключение

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.

Глава 3. Структуриране на текстови описания в биомедицината

Пациентските записи са богат източник на информация относно здравното състояние на пациента и лечението на неговото заболяване през годините, но те често съществуват само във формата на свободен текст. През 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* (Reuse of Electronic Health Records).

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

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

Въведение

Следвайки тенденцията за повторно използване на пациентски записи, ние разработихме методи за извличане на симптоми на диабета от медицински епикризи на български език. В болничните практики в България понякога лабораторните данни в ПЗ са налични само в текста, където лесно могат да бъдат разчетени от човек, но за успешен автоматичен анализ 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. Извлечената информация е ценна, защото от една страна е 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.

Фокусираме се върху извличане на *нива на кръвна захар*, *промяна в телесното тегло* и *полидипсо-полиуричен синдром*, които са едни от доминиращите фактори при поставяне на диагноза за диабет, както в върху *възрастта на пациента*. В примерите, които следват са показани изречения, в които се срещат тези симптоми, изказани по различни начини.

* 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**.
* Повод за настоящата хоспитализация е изтръпване и мравучкане на крайниците с намалена чувствителност и **лош кръвно-захарен контрол**.
* Постъпва по повод на **полидипсо-полиуричен 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**.

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

Дефиниция на задачата

Стремим да идентифицираме автоматично описанията на състояния в анамнезата на ПЗ. Формално погледнато всяко такова описание е релация между наименованието на симптома, неговия статус (посока на отклонение) и количественото му измерение. В изречението "При изследване **кръвната захар е била - 14 mmol/l.**" описанието се изразява в релацията между наименованието - *кръвна захар* и стойността на измерването - *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 - нала-

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. Оценяване на фаза 1. Резултати от подходите базирани на правила и на МС.

	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 4. Резултати от отделните фази на прилагане на правилата

гане 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.

Оценяване

Решаваме задачата за класифициране на изречения (Фаза 1) по два подхода - с правила и с машинно самообучение. 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 [37]. За класификацията използваме Weka Data Mining Software [16]. Най-добрите резултати са показани на Таблица 3 .

Таблица 3 показва, че точността на подходите, които са базирани на правила е по-висока от получената чрез класификация с машинно самообучение. Но при анализа на грешките забелязахме, че при правилата някои положителни примери са били погрешно класифицирани като положителни, защото въпреки че съдържат термини, описващи търсеното описание, съвпадението на правилата е станало с термини, които не описват търсените симптоми, а с други. 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 външно знание за областта, особено в задачи като тази, където точността на извличане е по-важна от покриваемостта.

Фазите 2-6 за извличане на симптоми и техните характеристики са базирани на правила. Резултатите от отделните фази на анализа са показани на Таблица 4 .

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%.

Резюме

Предлагаме унифициран подход за разпознаване на медицински описания на състояния, които имат смисъла на релации между няколко обекта - състояние, 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 състояния 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 (свеждане до корена на думата - стемиране) на изходните документи и ръчно добавяне на вариативност в правилата, по преценка на експерта. Разпознаване на класа на изреченията е направено с близка точност при двата подхода - бинарен J48 класификатор и с правила при анализа на *Фаза 1*. Тези резултати подсказват, че в бъдеще за подобни задачи е подходящо ползването на автоматична класификация с машинно самообучение, поради гъвкавостта на подхода и сравнително малките усилия за анотиране на корпуса на ниво изречние.

3.2. Разпознаване на събития и тяхната темпоралност в медицински текстове

Въведение

В областта на OEE се отделя сериозно внимание на изследване на времевия аспект. Когато извличаме връзки между обектите в текста, те често са валидна 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 зависи-

мост чрез следния пример: да се направи автоматично разграничение между споменавания на минали събития (напр. *Пациента 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 X, 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.

"Събитията" в медицинската терминология са широка категория с разнообразни дефиниции. В предишни i2b2 състезания[5], подобни задачи включваха извличане на медицински понятия, разпознаване на лекарства, разпознаване на състоянието на пациента и др., макар и с по-просто представяне[6]. Според [34], в състезанието i2b2 през 2010, системата, която е с най-добри резултати при извличането на понятия, използва условни случайни полета (CRF [23]). В някои от подходите авторите разделят проблема на подзадачи като *(i)* определяне границите на контекста и *(ii)* определяне на класа на извлеченото понятие [29]. Други групи обучават CRF модели с текстови характеристики, които са разширени с изхода от система, базирана на правила за разпознаване на именовани единици [14]. Други като [19] и [21] използват CRF модели, комбинирани със съществуващи системи за разпознаване на именовани единици и чънкери или подобни алгоритми, базирани на изхода от ресурси, обогатени с външно знание. 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 през 2010 [7] постига 0.852 F-мярка. Тя използва алгоритъм, подобен на CRF - дискриминативен полу-Марков HMM, обучен, използвайки пасивно-агресивно онлайн обновяване. Авторите извличат голям брой характеристики, между които са низове (tokens), конктекст, изречение, секция, характеристики на документа, характеристики извлечени от външни програми като cTakes[7] [13], MetaMap[8] [1] [2], ConText[9] [17], статистически парсери и др.

Авторът участва в колектив, представил система за състезанието 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

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

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

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

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

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

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.

Ресурси

Това изследване е проведено над данни, предоставени от Състезанието i2b2 2012[10], 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: (*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.

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

- повърхнинни;
- базирани на речници;
- контекстни;
- лексикални и морфо-синтактични;
- семантични.

Тези характеристики са извлечени за всеки низ (token) от учебните документи. Пунктуацията е изключена.

Експерименти и резултати

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 всеки низ, но решението за това, кои етикети формират едно събитие и кои ще са неговите атрибути, се взема на базата на правила. Като резултат бяха разработени 4 системи: базовата система, системата, чиито резултати предадохме на организаторите и 2 системи, които усъвършенствахме след това.

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

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, основа (stem).

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 низове, които представляват събитие, е различна.

Резултатите от четирите системи са описани в Таблица 5 .

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 5. Оценяване на извличането на събития върху тестовото множество от данни с програми, предоставени от i2b2.

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 за обработка на естествен език в биомедицината. 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, които 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 3. Интеграция в прототипи

3.1. Вграждане на машинно самообучение за класификация на атрибути в среда за автоматично откриване на потенциални диабетици в хранилище с големи данни

Въведение

Методите, които разработваме за извличане на информация от медицински документи, са успешно интегрирани и в система за анализ на амбулаторни листове (искания за финансиране, изпратени в НЗОК). Системата работи с милиони амбулаторни листове на пациенти 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 (изолирани) величини от голям обем текстове. Комбинацията от тези инструменти помага да създадем Регистър на диабета в България, а в представения тук експеримент ни дава възможност да открием потенциални диабетици, които не са формално диагностицирани с диабет. Техниките за ОЕЕ обработват само секциите за статус, лечение и анамнезата. Приносът на автора на дисертацията е в разпознаването на потенциални диабетно-болни пациенти.

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 листове в секцията *Диагноза*. 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 ОЕЕ, ползваме специфични характеристики и сигнализиращи думи. Авторът на дисертацията е работил по разпознаване на фрази, сигнализиращи диабет, в случай (*iii*). За извличането на тези събития направихме конкордансер и резултатите от него са обект на по-нататъчна обработка с ОЕЕ модули.

Дефиниция на задачата

За да се потвърди или отхвърли хипотезата, че текущият пациент "има диабет", прилагаме методи за машинно самообучение с учител за филтриране на текстовите чънкове от секциите със свободен текст в амбулаторните листове. За момента експериментираме с конкорданс на думата "*диабет*" в прозорец с 6 позиции (наляво и надясно). 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 ОЕЕ 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".

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, 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 разделен на словоформи, сведени до корен. Нашата задача е да класифицираме тези фрагменти, като потвърдим или отхвърлим хипотезата "има диабет". Това са няколко фрагмента, които демонстрират разнообразието от позитивни и негативни примери спрямо нашата хипотеза.

NEG Фамилност - обременен/а-**диабет**ици по майчина линия/.
NEG Необходимо е изключване на стероиден **диабет**; насочва се към ТЕЛК.
POS Покачва кръвно налягане. Има **диабет**. Оплаква се от сърцебиене...

Експерименти и резултати

За решаване 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*) Обучаване на модел чрез машинно самообучение за автоматично класифициране на записите от редуцираното множество на позитивни и негативни спрямо хипотезата. Експериментираме върху данните с 4 различни множества от характеристики:

- Експеримент 1. Булеви вектори от ръчно определени характеристики, извлечени само от положителните примери;

- Експеримент 2. Булеви вектори от ръчно определени характеристики, извлечени от положителни и отрицателни примери;

- Експерименти 3а и 3б. Булеви вектори от автоматично подбрани ха-

рактеристики;

- Експерименти 4а и 4б. Номинални вектори от автоматично подбрани характеристики.

Към тези множества от атрибути прибавихме и биграми и триграми. Резултатите от експериментите са показани на Таблици 6, 7 и 8.

	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
Позит.	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 6. Резултати от Експеримент 1 и 2 с ръчно подбрани характеристики.

	Експ. 3а: 151 авт. подбрани атр.: корени + бигр.; 10-fold cross-validation						Експ. 3b: 151 авт. 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
Позит.	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 7. Резултати от експерименти 3а и 3b - автоматичен подбор на характеристики, ползване на биграми и триграми.

	Експ. 4а: 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 8. Експерименти 4а и 4b с MaxEnt и всички текстови характеристики, биграми и триграми

Резюме

Този резултат представя първи стъпки към създаването на компютърна платформа за работа с голям обем медицински данни в 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 записи 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 величини и събития, свързани с лечението; класифицираме записите спрямо хипотезата "има диабет" с точност 91.5% и предлагаме тези наблюдения на специалистите, които вземат решения с цел подобряване на мениджмънта на българската здравна система.

Авторска справка

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 извлечен от първичните документи на експерименталния корпус. Технологията е разработена над 368 записа, извлечени от 1000 епикризи. 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 Int. 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: Int. 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.
http://www.lrec-conf.org/proceedings/lrec2014/pdf/675_Paper.pdf

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 конфенции:

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 по FP7 Capacity Programme на Европейската Комисия през 2012–2016.

Библиография

[1] Alan R Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proceedings of AMIA Symposium*, pages 17–21, 2001.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

[11] C. Friedman, P.O. Alderson, J.H. Austin, J.J. Cimino, and S.B. Johnson. A General Natural-Language Text Processor for Clinical Radiology. *JAMIA*, Mar-Apr, 1(2):161–74, 1994.

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

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

[14] 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.

[15] U. Hahn, M. 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

[21] 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.

[22] 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.

[23] 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.

[24] 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.

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

[26] Zubair Afzal Erik M van Mulligen Ning Kang, Bharat Singh 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.

[27] 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.

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

[29] K Roberts, B Rink, and S Harabagiu. 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.

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

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

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