Резюме

на публикациите на д-р Светла Николова Бойчева
представени за участие в конкурс за академична длъжност „доцент“ по 01.01.12 Информатика, за нуждите на ИИКТ-БАН, обявен в „Държавен вестник“ бр. 44 от 10/06/2016 г., стр. 99-100

Изискване	20 научни публикации	15 в списания с импакт фактор или в специализирани международни издания	20 цитирания	7 от цитиранията да са в списания с импакт фактор или специализирани международни издания	0
Изпълнение	22	22	86	59	1 защитил докторант

Представен е подход за извличане на информация от големи данни (37,9 милиона) амбулаторни листи на български език. Документите са в полуструктуриран формат – XML с големи полета текст в свободен формат съдържащ описанието на анамнезата, статуса, клиничните изследвания и лечението на пациентите. Целта на изследването е да се извлече информация за рисковите фактори за пациенти и да се покаже как могат да се интегрират констатациите в система, която дава възможност за ефективна превенция на диабет. Извлечени са данни от текстовите полета за клиничните резултати, статус и лечението на пациентите. Изследвани са различни комбинации от 2-6 рискови фактора за пациент. Резултатите показват слабите места в организацията на първичната медицинска помощ и специализирана извънболнична помощ в България. Резултатите са публикувани в [P1].

Направен е анализ на медицинската терминология, използвана в клиничните записи на български език. Един от основните проблеми е многоезиковата медицинска терминология – на български и латински език, както и използването на транслитерирани латински термини на кирилица. Друг основен проблем за автоматичната обработка на медицински текстове на български език е липсата на ресурси. Представен е подход за полуавтоматично създаване на българо-латински речник за медицински термини с около 5,000 единици. Като корпус са използвани класически медицински текстове на латински език и международни терминологични стандарти. Използвани са правила за превод на семантиката на латински медицински термини, в зависимост от комбинираните гръцко-латински корени, представки

и наставки и правилата за образуване на словоформи. Резултатите са публикувани в [P2] и е забелязано 1 цитиране.

Представен е подход за извличане на информация структурирана информация във вид атрибт-стойност за състоянието на пациент от епикризи на български език представени в текстов формат. Специфичната медицинска терминология на български и латински език и липсата на ресурси в електронен формат са някои от предизвикателствата. Използвани са статистически медоти за търсене на колокации, N-грами и словосъчетания, като са приложени различни методи Log Likehood, Mutual Information, $\chi - squares$. Дефинирани са правила и метрики за определяне на двойките атрибут-стойност за статуса на пациент. Представени са експерименти за обучаваща извадка от 1,300 и тестов корпус от 6,200 епикризи на български език за пациенти с ендокринни и метаболитни заболявания. При експериментите са използвани различни разстояния между думите в шаблоните вариращи от 0 до 50. Точността е около 96%, покриваемостта есравнително по-ниска 87%, поради множеството единични срещания на думи и използваните съкращения. Представеният подход е езиково независим и е тестван и за малък корпус на английски език с относително висока точност. Резултатите са публикувани в [P3] и са забелязани 3 цитирания.

Проектиран е метод за автоматично асоцииране на кодове по международната класификация на болестите – 10-та ревизия (ICD-10 – International Classification of Diseases 10th revision) на диагнозите в епикризи на български език в текстов формат. Описанието на диагнозите в епикризи съществено се различава от класовете диагнози в ICD-10 – използват се парафрази, описания на латински език, съкращения, както и спецификации. Предложеният подход е базиран на Support Vector Machines (SVM) метод, като всеки код от 4 значните рубрики на ICD-10 се разглежда като отделен клас. Разработен е прототип на система за автоматично асоцииране на ICD-10 кодове на диагнози представени като текст. Направени са експерименти с тестов корпус от 6,200 епикризи на български език за пациенти с ендокринни и метаболитни заболявания. Идентифицирани са описанията на 26,826 диагнози, за които са асоциирани 448 различни класове. Точността на асоцииране на кодове е 97.3%, покриваемостта е 74.68% и F1-measure е 84.5%. Основният проблем са множеството описания на латински език и използването на разнообразни съкращения. В 1,200 случая са използвани описания на процедури, които нямат съответствие по ICD-10, но за които има съответствие при предходната версия на класификацията ICD-9 и би било невъзможно дори и ръчно от

специалист да бъдат класифицирани. Резултатите са публикувани в [P4] и са забелязани 5 цитирания.

Представен е подход за идентифициране на темпорални маркери и сегментиране на историята на заболяването на пациент на епизоди. Разглежда се дискурса в описанието на секцията Анамнеза на епикризи на български език. Разработен е прототип на система за обработка на темпорална информация в медицински записи. Представени са експерименти с 1,375 епикризи на пациенти с ендокринни и метаболитни заболявания. Извлечени са 32,445 ключови термини от корпуса, в това число 7,000 срещания на имена на лекарствени продукти и около 7,500 срещания на наименования на диагнози. Темпоралните маркери се срещат 8,248 пъти обикновено съчетани с токени за времето идентифициращи посоката „преди“ или „след“ дадено събитие. Темпоралните маркери са идентифицирани с относително висока точност - 84%, но не много висока покриваемост. Резултатите са сравними с постиженията на подобни алгоритми прилагани за английски език, където се използват и много допълнителни ресурси. Резултатите са публикувани в [P5].

Анализиран е процеса на създаване на дизайн за обучение на учители в областта на технологиите. Идентифицирани са ключови компоненти, които са взаимосвързани с приложението на технологиите в процеса на професионална квалификация на учители. Анализирани са различни формални методи за представяне на знанията и е подбран формализъм за описание на модел на система за вземане на решение за дизайн на обучение на учители в областта на технологиите. Създадена е теоретичната рамка на модела, изградена на базата на изследването на експертното мнение. Представени са основните компоненти на модела. Описани са в термините на размитата логика под формата на лингвистични променливи и техните стойности характеристиките на компонентите на модела. Представени са взаимовръзките между променливите. Дефинирани са правилата, по които могат да бъдат правени изводи. Създадена е експертна система за подпомагане вземането на решение, базирана на модела. Приложимостта на модела е тествана чрез данните от четири различни дизайна на обучения. Резултатите са публикувани в [P6, P7, P8] Съавтор на статиите е Елиза Стефанова – докторант на С. Бойчева в периода 2009-2012 г.

Представени са подходи за структурирано извличане на информация за статуса на пациент от епикризи на български език в текстов формат. Използван е подход на плитък семантичен анализ. Създаден е прототип на система, която от описания в текста генерирани структурирани шаблони за текущото състояние на пациент. Направени са експерименти за

извличане на описанието на диагнозите, състоянието на крайниците на пациент, шия, щитовидна жлеза, възраст, пол, ИТМ, давност на диабет и др. Резултатите от експериментите показват висока точност при извличане на тези показатели. Създадени са алгоритми за нормиране на стойности на показателите в категории – добро, леко увредено, увредено и силно увредено. Резултатите са публикувани в [P9, P10, P11], като за [P10] са забелязани 3 цитирания.

Разработен е смесен подход за обработка на някои видове отрицание в изречения на естествен език, комбиниращ някои техники при извличане на информация и дълбок семантичен анализ на специализирани текстове. Реализиран е смесеният подход за обработка на отрицанието в система MEHR, който се използва за автоматично разпознаване и анализ на отрицанието в електронни записи на епикризи на пациенти представени на български език. Реализираната системата MEHR е предназначена за автоматична обработка на електронни записи на епикризи на пациенти, която автоматично попълва база с данни за здравния статус на пациента. Това е първото задълбочено изследване за автоматично извличане на информация медицински текстове на български език. Резултатите са публикувани в [P12], като са забелязани 13 цитирания.

Проектиран е алгоритъм за умозаключения при разбиране на парафрази чрез използване на знания за предметната област. Разработен е алгоритъм за «разбиране» на естествен език чрез умозаключения на базата на частичен семантичен анализ и обработка на темпоралната структура на текста. В този алгоритъм съществена роля играят дефинираните зависимости между събитията, които трабва да бъдат разпознати в текста. Разработеният алгоритъм при анализа обработва и някои видове отрицание и модалности. Реализиран е алгоритъмът за умозаключения при разбиране на парафрази и е интегриран в система FRET за извличане на информация от репортажи с цел попълване на шаблони за разпознаване на дадено събитие в тях. Резултатите са публикувани в [P13, P18]. Публикация [P13] има 1 цитиране, а [P18] има 12 цитирания. Публикация [P18] е препоръчана като основна литература в курс по Извличане на информация, Heinrich-Heine Universität Düsseldorf, Institut für Sprache und Information.

Проектиран е алгоритъм за определяне на семантичната коректност на изречение на естествен език, чрез проверка на неговата релация с минималното и максималното множество от допустими коректни изречения. Възможните резултати от анализа на изреченията са: коректно, частично коректно, непълно, грешно. При анализа се разпознава и

използването на парафрази. Анализът може да се прилага както върху пълно изречение, така и върху отделни фрази, които обикновено могат да се получават при т. нар. елиптичен отговор. Алгоритъмът може да обработва както отделни изречения или фрази, така и отговори, които съдържат няколко изречения. В последния случай се отчита и дискурса в параграфа. Реализиран е проектираният алгоритъм за доказване на семантична коректност на изречение и е интегриран в адаптивната система за електронно обучение STYLE за изучаване на чуждоезикова терминология в областта на финансовите пазари. Модулът за определяне на семантичната коректност на изречение се използва с цел проверка на коректността на отговорите на обучаемия, представени на естествен език при тестване на неговите знания. Реализираната система STYLE представя как при интегрирането на различни езикови технологии може да бъде реализиран адаптивен подход за електронно обучение. Използването на този модул в STYLE дава свобода на обучаемите да представят на естествен език своите отговори на задаваните им въпроси от системата, което позволява да се нарави по-задълбочен анализ от системата на знанията им. Резултатите са публикувани в [P14, P15, P16, P22]. Като съответно са забелязани следните цитирания: за [P14] – 3 цитирания, за [P15] – 1 цитиране, за [P16] – 6 цитирания и за [P22] – 8 цитирания.

Разработен е подход за автоматично генериране на въпроси на естествен език от учебни материали на базата на автоматично генериране на концептуални графи и чрез използване на онтологии представящи научната терминология в учебния материал. Изреченията в учебния материал трябва да отговарят на определени ограничения по отношение на тяхната синтактична структура. Представеният подход позволява до известна степен автоматизиране и подпомагане на процеса на създаване на експертния модул в системите за електронно обучение. На базата на проектирания подход за автоматично генериране на въпроси на естествен език от учебни материали е реализирана система CG-EST за електронно самообучение в специализирана област. Резултатите са публикувани в [P17] и са забелязани 4 цитирания.

Проектиран е алгоритъм за автоматичната обработка на някои видове отрицание във въпросителни изречения на естествен език в специализирани технически текстове в системи основани на знания. Разработеният алгоритъм позволява автоматичната обработка както на отрицателни, така и на положителни въпроси при някои ограничения на използваните езикови средства. На базата на проектирания алгоритъм е реализирана система основана на

знания за задаване на въпроси в областта на финансите. Знанията в ситемата са представени чрез концептуални графи. Резултатите са публикувани в [P19] и е забелязано 1 цитаране.

Разработен е подход за автоматично извличане на знания и генериране на концептуални графи от специализиран технически текст, представен на естествен език при някои ограничения. Генерираните концептуални графи се добавят автоматично в база от знания, съдържаща концептуални графи, като се проверява дали новопостроените графи, които ще бъдат добавени са консистентни с останалите графи, които се съдържат в базата. Реализиран е прототип на система CGExtract за автоматично извличане на знания и генериране на концептуални графи от икономически текстове представени на естествен език при някои ограничения. Генерираните концептуални графи се добавят автоматично в база от знания, съдържаща концептуални графи, като се проверява дали новопостроените графи, които ще бъдат добавени са консистентни с останалите графи, които се съдържат в базата. Резултатите са публикувани в [P20], като са забелязани 15 цитирания.

Разработен е подход за използване на концептуални графи за представяне на знанието в системи за електронно обучение. Акцентирано е върху проблемите при преподаване на чуждоезикова терминология. От една страна е необходимо да се изгради среда за електронно обучение, която поддържа представяне на знанието, което е интуитивно и удобно за визуализиране пред обучаемия. От друга страна трябва това формално представяне да позволява да се правят сложни анализи на коректността на отговорите на обучаемите, като се интегрират формални техники за разбиране на естествен език. Разработен е алгоритъм за трансформиране на концептуални графи във формално представяне чрез предикатна логика от първи ред. Проектираният алгоритъм е интегриран в система за електронно обучение в модул за проверка коректността на отговори на въпроси представени като свободен текст на естествен език. Резултатите са представени в [P21] и са забелязани 11 цитирания.