РЕЦЕНЗИЯ

от доц. д-р Иван Койчев, СУ "Св. Кл. Охридски"

на дисертационния труд на Ивелина Мирчева Николова,
на тема: „ПРИЛОЖЕНИЕ НА ОБРАБОТКАТА НА ЕСТЕСТВЕН ЕЗИК ЗА ИЗГРАЖДАНЕ НА СЕМАНТИЧНИ СИСТЕМИ"
представен за придобиване на образователната и научна степен „доктор"
в професионално направление 4.6 „Информатика и компютърни науки"

Представената дисертации е в много актуалната област - приложения на методите за обработка на естествен език за анализ на медицински текстове. Допълнително предизвикателство е, че са използвани текстовете на реални пациентски записи на български език, който е сравнително беден на компютърни езикови ресурси и инструменти с отворен код.

Дисертацията е добре структурирана. Състои се от увод, четири глави и заключение. Приложени са: списък на използваните термини, съкращения и означения; списък на таблиците; списък на фигурите; и списък с използваната литература.

Уводът дава кратка мотивация на темата, ясно определя целите на дисертацията и произтичащите от тях конкретни задачи.

Глава първа започва с представяне на обща архитектура на семантичните системи и разглежда техните основни компоненти. Прави обзор на съответните методите от областта на обработка на естествен език: разпознаване на парафрази, извличане на понятия и връзките между тях и автоматичното структуриране на описания от пациентски записи.

Глава втора разглежда приложение на методите за обработка на естествен език за създаване на модели на понятията за приложната област. Предложни са оригинални методи за разпознаването на синоними и парафрази и разпознаване на връзки между медицински понятия с използване на дефиниции от UMLS.

Глава трета разглежда задачата за структуриране на текстови описания в медицината. Предложени са оригинални подходи за откриване на синоними на диабет и разпознаване на събития и тяхната последователност в корпус от пациентски записи. Описват се експерименти с предложените подходи върху реални данни.

Глава четвърта представя внедряването на разработените методи за извличане на информация от медицински документи в система за анализ на амбулаторни документи. Използвани са и алгоритми за машинно самообучение за откриване на потенциални диабетици. Докладват се резултати от проведени експерименти с голямо множество данни от пациентски записи на български език, която цел е да се оцени точността на предсказване на системата.

Заключението обобщава постигнатите резултати и конкретните приноси на докторантката. Приложен е и списък с **литература**, съдържащ 67 заглавия, предимно научни статии и книги.

Основните приноси на докторантката, представени в дисертацията могат да бъдат обобщени както следва:

- Обзор на техниките за автоматично разпознаване на парафрази и синоними в компютърната лингвистика.
- Обзор на езиковите структури в пациентски записи на български език и избор на подходящи методи за тяхната обработка.
- Разработен е метод за автоматично разпознаване на парафрази като средство за справяне с разнообразието на изразяване на естествен език за нуждите на извличане на информация за състоянието на пациента.
- Реализиран е прототип за автоматично разпознаване на синоними и парафрази.
- Изследвани са медицинските онтологии, интегрирани в UMLS, относно наличието на явно представени връзки между понятията, като е разработен метод за използването им за развитие на понятиен модел за приложната областта на българския език.
- Разработен е метод за автоматично извличане на връзки между понятия от текст за развитие на понятиен модел на предметната област.
- Реализиран е прототип за извличане на връзки между понятията за медицински приложения.
- Планиране и провеждане на експерименти както с основани на правила подходи, така и със статистически методи за обработка на естествен език.
- Разработен е хибриден подход, при който статистическите техники подпомагат извличане на правила за анализ от текста на първичните документи от приложната област.
- Изследвани са техники за извличане на информация, които се използват при автоматичен анализ на медицински текстове.
- Разработени са прототипни компоненти за структуриране на информация за състоянието на пациента спрямо няколко важни и типични характеристики на пациентите.
- Проведени са експерименти с подходи, базирани на експертни правила и такива използващи методи за машинно самообучение с цел да се изследва и сравнена тяхната приложимост за дадената задача.
- Разработените прототипи са интегрирани в система за приложната област, като тестването на приложимостта и успеваемостта са извършени над реални данни на български език.
- Разработените методи за извличане на информация от медицински документи са успешно **внедрени** в система за анализ на амбулаторни листове на български език. Разработените компоненти за извличане на информация са тествани в продължение на няколко години в различни проекти. Тези софтуерни инструменти са използвани и при създаването Регистър на диабета в България за откриване на потенциални диабетици, които не са формално диагностицирани като такива.

Докторантката демонстрира много задълбочени знания в научните области на дисертацията. Направените обзори са фокусирани и адекватно представят текущото състояние на изследванията. Списъка с използвана литература е представителен за областта и е достатъчен за докторска дисертация. Представянето на разработените подходи е добре структурирано и подкрепено с подходящи илюстриращи примери. Всяка секция завършва с обобщение на постигнатите резултати, като се посочват и възможности за продължаване на изследванията.

Избраната методологията на изследване включва: аналитичен обзор на областта, мотивиран избор на подходящи методи и тяхното творчески прилагане и доразвиване за дадената приложна област. Разработени методи са реализирани като софтуерни прототипи след което експериментално са изследвани и верифицирани с реални данни. Част от разработените компоненти са внедрени в реални проекти. Приложената изследователска методология е подходяща за поставените цели в дисертацията и е традиционен за научната област.

Много добро впечатление прави и оформянето на самият текст, който е добре структуриран и естетически оформен. Издържан е и от гледна точка на правопис, граматика и стилистика. Важните единици в текста са обозначени и почертани с подходящо форматиране, което ги прави откроявящи се от останалият текст. Фигурите и таблиците са надлежно номерирани, имат обяснителни надписи и са цитирани в текста.

Също така добро впечатление прави, че термините са надлежно преведени на български, като към дисертацията е приложен речник на термините.

Приносите на докторантката са значителни и определено са достатъчни за дисертационен труд за образователна и научна степен доктор. Публикувани са в 10 статии, като на 6 от тях докторантката е водещ автор. Всичките са публикувани в сборници от доклади на авторитетни международни конференции. Четири от тези сборници са издадени в серии на реномираното издателство Springer. Забелязани са 4 позоваванията на публикации по дисертацията и то в статии на реномирани конференции и научни списания. Приемането на статиите на специализирани рецензирани конференции в съответните области и цитиранията им са убедителни доказателства за тяхната значимост.

Авторефератът е добре направен, като вярно и точно отразява дисертационния труд.

Към представената дисертация нямам съществени забележки, все пак има няколко бележки, които е добре да бъдат споменати и евентуално разгледани при подготовката на окончателният текст:

- Може още да се подобри преводът на термините. Например: след като терминът concept се превежда като „понятие“, то терминът conceptual model може да се преведе като „понятиен модел“ и др.
- Хубаво е в глава първа да се дадат по-формални определения на по-важните понятията като: „парафраза“, „модификатор“, семантична структура и др.
- Въведената на страница 17 F-мярка по-често се бележи като F_1-мярка.
- Мярката Precision е преведена като „точност“, което е много добър превод, но се препокрива с установеният превод на термина Accuracy, който е различен, но често се използва като друга мярка за оценка на класификатори.

Тези бележки по-скоро адресират неутвърдената терминология в областта, особено на български, но това е нормално състояние за една сравнително млада и все още бурно развиваща се област. По-подробен списък с конкретни предложения за превода на някои от термините ще предоставя на докторанта.

Професионалната биография на докторантката е респектираща и говори, че тя е високо квалифициран специалист в областта на информатиката с богат професионален опит, натрупан при работа по редица научно-приложни проекти, много от които международни.

Можем да **обобщим**: от представеното дотук е видно, че представената работа има съществени научно-приложни приноси, напълно достатъчни за присъждане на научно-образователна степен „доктор”.

В заключение давам **положителна** оценка на представената дисертация и предлагам на уважаемото научно жури, по обявената процедура за защита на докторска дисертация в ИИКТ-БАН да даде образователната и научна степен „доктор” в професионално направление 4.6. „Информатика и компютърни науки“ на автора на дисертационния труд Ивелина Мирчева Николова.

Дата 29.12.2014 г.