РЕЦЕНЗИЯ

на дисертационния труд на Ивелина Мирчева Николова на тема
„Приложение на обработката на естествен език за изграждането на семантични системи"
за присъждане на образователната и научна степен „Доктор" в професионално
направление 4.6 „Информатика и компютърни науки"
от доц. д-р Кирил Иванов Симов, ИИКТ, БАН

Трудът на Ивелина Николова съдържа 129 стр., структурирани в Увод, 4 глави, Заключение, Списък на използваните термини, съкращения и означения, Списък на фигурите, Списък на таблиците и библиография. Ясно са посочени научните и научно-приложните приноси на дисертантката. Прилагането на списъци с използваните термини и на фигурите, както и таблици подпомагат четенето на дисертацията и подсилват аргументативната част.

Темата изследва възможностите на обработката на естествен език (ОЕЕ) за създаването на приложения, основаващи се на семантични технологии. Изследването е ценно в поне няколко посоки: показва важността от наличието на разнообразни и богати ресурси за естествените езици за създаването на полезни приложения; показва начините, по които ОЕЕ подпомага семантични архитектури -- заедно с предимствата и ограниченията; доказва се, че съществуващите ОЕЕ дават добри резултати в конкретни предметни области. Представени са приложения на ОЕЕ за създаването както на терминологични семантични ресурси (онтологии) на базата на документи от предметната област, така и за извличане на конкретни факти за обекти описани в тези документи. Фактите представляват конкретни обекти, техни характеристики, релации между тях, събития и техните времеви параметри. Така дисертацията представя разработки на методи и приложения за анализ на един цял диапазон от явления, необходими за успешни семантични приложения.

Дисертантката познава отлично литературата по темата, умее да мотивира и аргументира решенията си, както и да структурира много добре съдържанието на текста.

В Увода се разисква актуалността на темата и целите и задачите на дисертационния труд. Мотивирана е нуждата от интегриране на системите за обработка на естествения език със системите, базирани на семантични технологии. Поставени са основните цели пред една такава интеграция. Тези основни цели са развити в конкретни цели и задачи, които да бъдат решени в рамките на дисертационния труд. Уводът завършва с представяне на структурата на дисертацията.

Глава 1 „Обзор на основните резултати в областта" представя детайлно описание на характеристиките и архитектурата на семантичните системи. В рамките на описанието на семантичните системи се представя детайлно и мястото на ОЕЕ в рамките на една семантична система. Останалата част от главата съдържа обзор на основните резултати в областта. Резултатите са групирани според основните задачи на дисертацията:

разпознаването на парафрази; извличането на понятия и релации; автоматичното структуриране на описания от пациентски записи.

Глава 2 „Приложение на OEE за създаване на концептуални модели на предметна област" се съсредоточава върху два подхода, които подпомагат построяването на модел на определена предметна област: (а) разпознаване на парафрази на понятия и (б) обогатяването им с релации. Подходите са демонстрирани в медицинската област, и по-конкретно – в сферата на болните от диабет в български контекст. Разгледани са релации от типа IS-A и AFFECTS.

Дисертантката предлага идея за създаване на модел на данни от ресурсно небогат език, какъвто е българският, като използва трансфер на знание от английски ресурси (като UMLS - Unified Medical Language System) в български текстове и прилага обработки върху тези текстове (токанайзер и стемер). Данните, с които се работи, са около 1000 структурирани епикризи от български болници с дължина 2-3 страници. След това се решават задачите за разпознаване на синоними и парафрази и обогатяване с релации между понятията. Първоначално се прилага статистически анализ върху наличните текстове, за да се отсее значимата терминология. Извлича се терминологичен речник и речник с колокации на термините, които се подават към UMLS, а после се връщат обратно чрез превод на намерената информация на български. Приносен момент е извеждането на правила за близост на понятия (положителни и отрицателни). Дискутирани са проблемите на многозначността и адаптирането към конкретните предметни подобласти; взета е предвид и ролята на контекста. Показано е, че правилата могат да се използват за други езици и за други предметни подобласти. Направена е оценка на предложения подход върху 368 описания на крайници с речник от 265 термина. Резултатите са много добри – 96% точност и 89% покриваемост. Разбира се, трябва да се отчете фактът, че речникът е малък и е бил преведен ръчно на английски. Това означава, че когато се работи с качествени ресурси, резултатите са високи. Както и самата дисертантка подчертава, интересно ще бъде да се види резултатът при въвеждане на компонент за автоматичен превод, който със сигурност би внесъл шум в данните чрез многозначността и неточността. Това повдига и въпросът как е бил преведен речникът за споменатите експерименти? Дали той е преведен от професионален преводач или от експерт в предметната област и двете може да имат и положително и отрицателно влияние върху изпълнението на системата.

За извеждане на релациите отново се разчита на UMLS, откъдето се извличат дефинициите на термините, които отговарят на българските термини. Използва се и програмата RelEx за извличане на депендентни структури. Стъпките изискват филтриране на входните данни UMLS по отношение на последващите обработки: изчистване на редундантна информация и трансформиране на безглаголни структури в глаголни.

Представени са правила за релациите IS-A и AFFECTS. Първата релация се реализира чрез 2 правила, параметризирани спрямо особеностите на българските текстове. Извличането ѝ е сравнително тривиално, но изключително важно за модела. По-сложно се обработва втората релация, тъй като тя се извлича от текста, а не от структурирани данни като дефинициите. Като предимство отчитам прилагането на лингвистичен подход при

създаването на правилата за извличане на релации. Разбира се, всяка последваща обработка зависи от резултата на предишната – в случая – депендентния парсер.

Глава 3 „Структуриране на текстови описания в биомедицината" се фокусира върху обработката на текстове в областта на медицината с цел извличане на полезна информация от тях. По-конкретно, дисертантката разглежда извличането на симптоми на диабет от български пациентски записи. Приносът на тази глава е във факта, че дисертантката представя свой метод за обработка на данните, който съвместява техники на машинното самообучение и анализи чрез лингвистични правила. Освен това, приложимостта на подобно средство е от изключителна важност, тъй като то би подпомогнало лекарите при вземането на решение за болния чрез агрегиране на подходящи данни и спестяване на време за преглеждането им.

Ивелина Николова избира симптомите 'нива на кръвна захар' и 'промяна в телесното тегло', за да докаже адекватността на предложения подход. Направен е анализ на особеностите на текстовете. Това е необходима стъпка, за да се адаптират средствата за обработка в необходимите посоки. Алгоритъмът работи с множество речници, които отчитат специфичните термини и контекста. Модулът с правила е построен на принципа top-down – от изречението към симптома и неговия статус и т.н. Те са 6 на брой. Всички подмодули са описани поотделно. Повечето от тях са базирани на правила, а само при класифициране на изреченията със симптоми се използва и машинно самообучение. При него са направени експерименти с различни алгоритми, като най-добре представилият се е J48. Оценката показва, че правилата са по-добри от машинното обучение, но въпреки това то също дава добри резултати. Даден е подробен анализ на проблемите при прилагане на отделните стъпки.

Във втората част на тази глава се представя система (няколко различни версии според използваните ресурси и подходи) за разпознаване на събития и времеви етикети в медицински текстове. Тук дисертантката представя авторската си разработка за разпознаване на събития чрез метод за машинно самообучение с учител и с правила върху медицински данни на английски език, проведени в рамките на състезание за темпорални системи. Използвани са средствата MetaMap за извличане на семантична информация и GATE за текстова обработка. Използван е алгоритъмът CRF с 5 типа характеристики (повърхнинни, базирани на речници, контекстни, лексикални и морфосинтактични, семантични). Сравнени са 4 системи, създадени от дисертантката – една базова и три подобрени. Последната система отчита и последователността на събитията. Тя е и най-добрата като резултати. Системата е разработена за английски с цел участие в състезанието i2b2 2012. Това участие и получените резултати е гаранция за качеството на разработената система. Тъй като се разчита на семантични екстрактори като MetaMap и подходящо анотирани медицински текстове, остава въпросът как може подобно средство за извличане на събития да се локализира за български данни? Колко голямо усилие е да се създадат необходимите ресурси за целта? Има ли възможност за подобрение в съответните ресурси?

Глава 4 „Интеграция в прототипи" описва вграждането на компоненти за извличане на знание в конкретни софтуерни архитектури. Тази глава представя интеграцията на описаните в предходната глава методи за извличане на информация от медицински текстове в система за анализ на амбулаторни документи. Тук подробно е представен компонент, който разпознава болни от диабет хора чрез тестване на верността на хипотезата 'има диабет'. Като входни данни са използвани конкордансите на думата 'диабет'. Отново е приложен хибриден подход, което според мен е много добра стратегия . Правилата са използвани за филтриране. Целта е да се получи висока точност при позитивните примери, за да се идентифицират нови болни. Представени са 4 експеримента, като във всеки броят на характеристиките нараства, което води до по-голяма точност (91.5), но покритието остава малко (22.6). По-важна е точността, но както е отбелязано от самата дисертантка, ролята на тестовите данни като качество и количество е значителна, тъй като е необходим голям брой положителни данни. Въпреки малките данни обаче експериментите показват добра приложимост за този домейн. Очакването е резултатите да се подобрят чувствително при големи данни, ако те станат налични.

Заключението обобщава съдържанието и резултатите на дисертацията, както и бъдещите насоки на работа. Публикациите на дисертантката по темата са 10, като всички са в реферирани издания или в сборници на престижни международни конференции. Има и 4 отбелязани цитирания на част от статиите. Основните научни и научно-приложни приноси са посочени коректно. Резултатите са получени и използвани в размките на няколко национални и международни проекта, което е допълнителна гаранция за качеството им.

Авторефератът отразява коректно целите, задачите и резултатите на дисертационния труд. Малка забележка към него е, че е объркана номерацията на глава 4 и тя е номерирана като глава 3.

Смятам, че дисертацията на Ивелина Николова има всички качества, изисквани за дисертационен труд. Дисертантката има редица приносни моменти в решаването на задачи, свързани с обработката на текстове в медицинския домейн и с преноса на знание от език с повече към език с по-малко ресурси. Един въпрос за всички използвани методи е въпросът за използване на по-голям контекст от едно изречение. Този въпрос е по-скоро за бъдещата работа по темата и не влияе върху качеството на дисертацията.

Убедено препоръчвам на уважаемото жури присъждане на образователната и научна степен „доктор" на Ивелина Мирчева Николова.

30 декември 2014 г.
София