Институт по информационни и комуникационни технологии-БАН
Вх. № 2 / 05.01.2015 г.

СТАНОВИЩЕ

за дисертационния труд на Ивелина Мирчева Николова

ПРИЛОЖЕНИЕ НА ОБРАБОТКАТА НА ЕСТЕСТВЕН ЕЗИК ЗА ИЗГРАЖДАНЕ НА СЕМАНТИЧНИ СИСТЕМИ

от доц. д-р Мария Дечова Стамболиева, НБУ.

Дисертационният труд на Ивелина Мирчева Николова е посветен на една безспорно актуална тема, каквато е модернизацията на българското здравеопазване и подобряването на управлението на здравната ни система. Целта на работата – създаването на ресурси и софтуер за обработка на медицински записи – ясно говори за интереса на докторантката към значимите проблеми на съвременния ни живот. За постигане на така поставената цел, Ивелина Николова формулира следните задачи за разрешаване: 1/изследване на техниките за автоматично разпознаване на парафрази в компютърната лингвистика и разработване на прототипи; 2/ разучаване на съществуващия софтуер за извличане на релации между понятия от големи медицински онтологии; 3/ реализиране на прототип за извличане на релации; 4/ разработване на хибриден модел за извличане на релации и разпозвнаване на парафрази; 5/разработване на прототипни компоненти за структуриране на информацията за състоянието на пациента; 6/ интегрирането на така създадените прототипи в системи за автоматична обработка на естествен език. Дисертацията е структурирана в четири глави, следвани от списък със статии на докторантката по темата на дисертацията, апробация на резултатите, списък на съкращения и библиография.

Първата глава е сериозен обзор на постигнатото в областта, основан на анализа на десетки трудове, който обзор демонстрира отличната теоретична подготовка на докторантката и познаването на подходите, изследванията и разработките в областта на автоматичното разпознаване на парафрази, извличането на понятия и релации и автоматичното

структуриране на описания от пациентски записи. В този обзор докторантката обръща внимание и на отделни слабости на съществуващите прототипи, като например недостатъчното внимание, което се отделя на обработката на отрицанието. Техническа грешка вероятно е твърдението, че „През 1959 г. Сосюр дава едни от първите класически определения на релации между понятията" (с. 18) – докторантката е имала предвид датата на издаване на цитирания труд.

За разработването на приложението за създаване на концептуални модели на предметната област, на което е посветена втората глава, са използвани пациентски записи с медицинска терминология, която описва заболяването диабет. Вниманието на И. Николова е насочено към две основни връзки, IS-A и AFFECTS, чието наличие позволява правенето на обобщения върху извличаните от документите факти – една възможност, която авторката добре е илюстрирала с описанието на състоянието „онихомикоза". Сериозен принос е както разработената методика за попълване на оскъдните български ресурси чрез запитвания към съществуващи англоезични или многоезични терминологични ресурси, така и извличането на нови релации. Макар че тази бележка няма отношение към крайния резултат на анализа в тази част, бих искала все пак да подчертая, че структурата NP IS A NP, макар и да следва словоредния модел S V O, не го илюстрира, тъй като втората именна фраза е предикатив, а не допълнение. Предикативът препраща към подложната фраза, като я характеризира и така, по един или друг начин, дефинира. Релацията AFFECTS се реализира в структурите NP V NP/PP и лесно може да се извлече от речник-лексикон с изброени семантични роли на аргументите към глаголи със съответна семантика. В тази част на работата, последователните стъпки в синтактико-семантичния анализ са много ясно изброени и обобщени в правило (2.3.2). Изненадваща е неспособността на използвания парсер да разреши многозначността СЪЩЕСТВИТЕЛНО/ГЛАГОЛ в английски език в ясните подсказващи контексти (с. 57): НЕОПРЕДЕЛИТЕЛЕН ЧЛЕН – JOINT -ГЛАГОЛНА ФОРМА или DET JOINT PP. Независимо от това, постигнатите резултати са окуражаващи и

подсказват възможности за приложение в нови задачи, като извличане на енциклопедична и справочна информация за учебни цели.

Безспорно приносно е и изследването на възможностите за повторно използване на пациентски записи чрез автоматично структуриране на свободен текст. Една бележка: ако в изследването методът „стеминг" се състои в изчистване на флексията (не ИНфлексията!), той оставя основа, а не корен. За да се стигне до корена, трябва да се отстранят и деривационните афикси. Докторантката обаче говори за корени, до които се стига чрез изчистване на словоизменителни афикси (флексии). И в тази глава прави добро впечатление ясното описание на последователните процедури на анализ (от фаза 1 до фаза 6). Малко разочароващо е описанието на обработката на отрицанието във фаза 5 (само 6 реда), тъй като това е един от приносите на дисертацията. Получените резултати обаче са отлични (вж. таблици 3.6.3., 3.7.4), очертани са добри перспективи за машинно самообучение и възможности за продължение на изследването. В последната част на главата, много добре са представени процедурите за разпознаване на събития и са очертани, макар и само с един пример, възможни подходи за установяване на темпоралност.

Четвъртата глава демонстрира възможностите за използване на разработените методи в системи за анализ на документи. Конкретната цел – разпознаване на пациенти, болни от диабет, които не са диагностицирани с това заболяване – е постигната и ясно демонстрирана с описанието на проведените експерименти и анализа на постигнатите резултати.

В обобщение:

Дисертацията на Ивелина Мирчева Николова е сериозен научен труд, плод на задълбочена проучвателна и изследователска работа, с ясно приложение в една важна за страната област, която се нуждае от иновации, и с добре очертани перспективи за надграждане. С поставените задачи, използваната методология,

богатството на ползвана литература и постигнатите резултати тази дисертация напълно отговоря на изискванията за присъждане на образователната и научна степен „Доктор” в научната специалност 01.01.12. „Информатика”.

София, 30.12. 2014