РЕЗЮМЕ

на научните постижения в статиите, представени от доц. дмн Галя Ангелова
при кандидатстване на конкурс за професор по специалността
Информатика (Езикови и семантични технологии), септември 2011 г.

За конкурса са представени 15 статии в чуждестранни рецензирани издания, публикувани след 2003 г. По-голяма част от статиите (11 на брой) са публикувани в периода 2009-2011 г. Статия № 1 е в издание на Шпрингер с JRC/ISI impact-factor 0,251. Статии № 11 и № 15 са в серията "Studies in Health Technology and Informatics" на IOS Press с H-index 21. От представените 15 статии, 14 са публикувани в списания или томове от научни поредици на реномирани издателства, а една статия - № 6 – в Сборник трудове на семинар, съпътстващ престижната международна конференция LREC (Language Resources and Evaluation Conference). Всички представени трудове не са използвани в предишни процедури за научни степени или звания.

Статиите съдържат резултати в три области:

- Взаимодействие между езикови и семантични технологии в семантичния интернет или при построяване на декларативни семантични ресурси и дигитализирани архиви от учебни обекти (3 статии);
- Изучаване на поведението или построяване на приложения чрез статистически методи за обработка на естествен език (2 статии) , и
- Автоматична обработка на български текст в клинични записи на пациенти (10 статии).

1. Взаимодействие между езикови и семантични технологии (статии № 1, 4 и 9).

Статия № 1 (публикувана през 2004 г.) представя тезата, че – независимо от желанието на редица изследователи и потребители да се построи интернет над знания (а не над текстове), в момента и в обозримо бъдеще основният носител на информация в интернет е текстът на естествен език. Макар че Семантичният интернет се построява над онтологии, които биха могли да се използват като метаданни и анотационни "котви" на единиците от текстовете, такива онтологии не са налични в достатъчно количество. Построяването на хипервръзки към онтология-скелет за анотация на понятията също не е тривиално, тъй като хипервръзка се установява от дума в текста към етикети на понятия, написани на същия естествен език като анотирания текст. Езиковите технологии, които по време на публикуването на статията се разглеждаха като обещаващ инструмент за автоматично създаване на онтологии чрез анализ на текста, предлагат само средства за натрупване на онтологични полу-фабрикати. В статията се предлага подход за полу-автоматично анотиране на икономически текстове

чрез прекарване на хипервръзки към понятията на онтология в областта на финансовите пазари, създадена в ИИКТ-БАН под ръководството на Г. Ангелова в проекта Ларфласт през 1998-2001 година. Чрез автоматичен анализ на текста се разпознават части от имена на понятия и се предлагат оригинални визуални средства за дефиниране на хипервръзки. Създадените анотирани текстове се използват като ресурс за семантично търсене и илюстрират предимствата на Семантичния интернет.

Статия № 4 (публикувана през 2007 г.) анализира възможностите за приложение на модела LCIM (Levels of Conceptual Interoperability Model)[1] към проблемите на семантичната интероперабилност на учебни ресурси в Семантичния интернет. Макар че LCIM оригинално е ориентиран към композирането на системи при моделиране и симулация, статия № 4 предлага използването на LCIM като методическа рамка при създаване на съвместими и взаимодействащи си учебни ресурси. Онтологиите се разглеждат като таксономии от понятия, чрез които се постига унифицирано смесване на учебни ресурси от два различни архива. Направен е експеримент за "смесване" на учебни обекти в банки с дигитални учебни обекти в областта на финансовите пазари. Това "смесване" се осъществява чрез усъвършенстване на метаданните към учебните обекти, като по този начин обучаемият не се натоварва с идентификация на оригиналния първоначален ресурс. Създадено е средство за визуализация на понятията от двата ресурса, което позволява на потребителя да управлява новокомпозирания ресурс от интероперабилно учебно съдържание. През 2008-2009 тази статия се цитира 4 пъти от Андреас Ток и Чарлз Турница, които са измежду основните автори, споменати на посочената страница на LCIM в Уикипедия.

Статия № 9 (публикувана през 2010 г.) анализира състоянието на езиковите технологии, които могат да извлекат автоматично "концептуални" единици от технически текст – като етикетът на тези единици е термин. Езиковите технологии са групирани като типове продукти, които откриват автоматично в текста *(i)* термините, състоящи се от няколко думи; *(ii)* връзката понятие-надпонятие (термин-обобщение) в полу-структуриран текст, *(iii)* отношения между понятия, които заемат подобни синтактични позиции в изреченията – и така може да се установи например, че *ваксинация, ваксина, доза, инжекция* са термини от един семантичен клъстер; *(iv)* връзката понятие-надпонятие (термин-обобщение) в свободен текст чрез прилагане на автоматично самообучение с цел построяване на таксономии от термини; в последната категория се включват и подходи за построяване на концептуални йерархии по модела на формалния концептуален анализ след автоматичен синтактичен разбор на изреченията в текста. Групирането на езиковите технологии е подкрепено от анализ на коректността, с която работят разгледаните приложения за анализ на естествения език. Обоснован е изводът, че езиковите технологии за обработка на думи и изречения имат готовност да се прилагат в софтуерни системи, произвеждащи сурогати на декларативни концептуални структури.

---

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

Съавтори на две от представените статии са Албена Струпчанска, Павлин Добрев и Огнян Калайджиев, които в съответните години на публикуване са били докторанти на кандидата.

2. Приложения на статистически методи при автоматичната обработка на естествен език (статии № 2 и 3)

Статия № 2 (публикувана през 2003 г. и в съкратен вариант през 2004 г.) представя експерименти, направени през 2002-2003 година с цел изучаване на поведението на т.нар. Latent Semantic Analysis (LSA), при който документите се представят като вектори от думи (bag of words) и оригиналното много голямо пространство се свива до по-малко по размер чрез метода на SVD (Singular Value Decomposition). Целта на експериментите е да се установи дали и доколко предварителният лингвистичен анализ влияе върху резултатите за намиране на близост между документите. Работи се със 702 различни документа на български език. Предположението е, че за българския – като силно флективен език с много словоформи – е от голямо значение разнообразните словоформи да се сведат до основната и текстът да се "нормализира", което би изменило размерността на оригиналното векторно пространство, съответстващо на разглеждания корпус. Думите в документите от тестовия корпус са представени по 6 начина: *(i)* оригинален графемен вид, *(ii)* думи заместени със stems след stemming за всяка словоформа, *(iii)* думи заместени с основни форми за всяка словоформа след морфологичен анализ, *(iv)* думи групирани във фразови единици за по-важните термини, *(v)* думи заместени с основни форми за всяка словоформа след морфологичен анализ и групирани във фразови единици за по-важните термини и *(vi)* думи, заместени с основна форма, с разрешена многозначност като част на речта. Като контролни методи за измерване на близост, независими от ЛСА, се използват *(i)* k-nearest neighbour classifier и *(ii)* класическият метод за търсене на косинус-близост във векторно пространство без свиване на размерността. Освен това, експериментите включват тестове със и без стоп-думи. Направена е серия от тестове, които показват, че най-важният фактор е изборът на схема за пресмятане на тегла, присвоени на локално ниво (тежест на термина в документа) и на глобално ниво (тежест на термина в целия корпус). Другите популярни параметри като напр. премахване на стоп-думите и свиване на размерността не оказват почти никакво значение за подобряване на резултатите. Стемуването и лингвистичният анализ се оказват почти еднакво добри за силно-флективния български език, което е обезкуражаващо за компютърната лингвистика, тъй като се оказва, че дефиницията на понятието «дума» няма значение при изчисляване на близостта между документите за този сравнително малък корпус. По-сложните лингвистични техники за предварителен анализ на текста (групиране във фрази, снемане на многозначността) не подобряват резултатите. ЛСА е по-добър от изчисленията в оригиналното векторно пространство, понеже при малкия експериментален корпус редукцията позволява събиране на сравнително редките

термини, разпръснати в оригиналното пространство по различни оси. Тази статия е една от малкото публикации, посветени на сравнителен анализ на различни техники за измерване на близост между документи. Цитирана е 11 пъти и е препоръчана в учебен курс в Германия. Съавтори на статията са Преслав Наков (през 2003 г. бивш дипломант на Г. Ангелова) и Елена Вълчанова (докторант на Г. Ангелова през 2003 г.)

Статия № 3 (публикувана през 2003 г. и в съкратен вариант през 2004 г.) представя системата MorphoClass за автоматична обработка на немски текст, която «предсказва» морфологичния клас на непознати немски думи, «приличащи» на съществителни имена в съответни позиции в текста. Системата беше развита над дипломната работа на Преслав Наков, защитена през 2002 г. под ръководството на Г. Ангелова. MorphoClass е прототип за извличане на лингвистични знания: той *(i)* намира непознати думи в текста (т.е. думи, които не са включени в речника му), *(ii)* установява свойствата им, *(iii)* групира ги с други словоформи - потенциални кандидати да бъдат словоформи на същата дума, *(iv)* разделя сложните немски съществителни на съставни части, ако срещне непозната дума от този вид, *(v)* генерира хипотези коя е основата на непознатата дума за всяка група намерени словоформи и *(vi)* класифицира непознатите думи в най-вероятния морфологичен клас. Чрез наблюдения на големи масиви от думи и използвайки знания за немската морфология на съществителните имена, системата създава множество от ending-guessing rules, които прилага при срещане на непознати думи с отчитане на тяхната вероятност. Локалният контекст на съседните думи също се взема под внимание при избиране на вероятния клас на непознатата дума. Системата е тествана над корпуси от произведения на Кафка и Гьоте с големина над 280000 словоформи. Постигнатата точност е над 74% при покриване на 89% от непознатите думи. Подходът на MorphoClass за разпознаване на непознати думи е приложен за исландски текст (вж. списъка от 11 цитирания) и изглежда се оценява като важен за езици, за които няма създадени големи електронни речници. Съавтори на статията са Преслав Наков (през 2003 г. бивш дипломант на Г. Ангелова) и Юри Бонев (през 2003 г. дипломант на Г. Ангелова), както и Евелин Гиус и Валтер фон Хан от Университета в Хамбург (които предоставиха много ресурси и участваха в оценяването на системата).

3. Автоматична обработка на български текст в клинични записи на пациенти (10 статии)

Статии № 5, 6, 7, 8, 10, 11, 12, 13, 14 и 15 представят резултати в областта на автоматичната текстообработка на медицински текстове, получени през 2009-2011 г.

Публикации № 5, 6 и 8 представят дизайна на една прототипна система, първи успехи при извличане на отделни стойности и при разработване на софтуерна среда за поддържка на натрупваните автоматично извлечени данни за пациенти. Статия № 5 (отпечатана през 2009 г.) представя дизайна на научния прототип ЕВТИМА, който е в процес на разработка под ръководството на Г. Ангелова в проекта Д0-02-292

"Ефективни методи за търсене на концептуални шаблони с приложения в медицинската информатика", финансиран от Фонд "Научни изследвания" през 2009-2012 г. Целта на прототипа е да търси концептуални шаблони в болнични записи на пациенти-диабетици. Като първоначална задача се поставя извличане на концептуални единици от текста на записите, като целта е единиците да се разпознаят в текста и да се структурират във вътрешно представяне от типа на семантична мрежа. В статия № 5 се анализират трудностите за автоматичен анализ на медицински текстове, измежду които са наличието на много латински термини и съкращения в епикризите, както и липсата на електронни ресурси на български език вкл. на терминологични речници. Представен е един основен принцип на дизайна на системата: за всеки пациент да се създаде един концептуален граф, който да бъде "налаган" към графите на други пациенти с цел търсене на подобие. Статия № 6 (публикувана през май 2010 г.) представя разработените към момента прототипи за извличане на структурирана информация за *статуса* на пациента, които извършват плитък синтактичен анализ на базата на регулярни изрази. Регулярните изрази са "научени" полуавтоматично от системата по метода на машинното самообучение чрез наблюдение на описания на статуса в голям корпус от болнични записи на диабетици. Представени са вътрешните шаблони, в които системата натрупва извлечените стойности на различни показатели. Резюмирани са резултатите от експерименталите оценки на прототипните процедури за извличане: постигната е точност над 90% при автоматично разпознаване на описания на диагнозата, състоянието на шията и щитовидната жлаза на пациента; точност над 80% при автоматично разпознаване на описания на възрастта, продължителността на диабета, и състоянието на крайниците; и точност 61% при автоматично разпознаване на пола на пациента (което е затруднено поради изтриването на лични данни при анонимизацията на записите). Статия № 8 (публикувана през септември 2010 г.) представя софтуерната среда ЕВТИМА, която поддържа натрупваната структурирана информация за отделните пациенти. Едно съществено улеснение при автоматичното извличане е наличието на зони в текста на епикризите, съгласно Наредби за съдържанието на медицинските документи в Република България, публикувани в "Държавен вестник". Макар че структурата не се спазва съвсем стриктно на практика, тя осигурява зонирането на епикризите с голяма точност, тъй като в текста на анонимизираните епикризи се откриват с почти 100% точност заглавия на най-важните зони (диагнози, анамнеза – история на заболяването, статус, данни от клинични изследвания, обсъждане, лечение, препоръки). В статия № 8 се представя и постигнатата точност на зониране на епикризите от изпозвания през 2010 г. корпус от 1197 болнични записа – тя е над 98%.

Публикации № 7, 10, 12 и 13 представят по-нататъшни резултати за обработка на концептуална информация в системата ЕВТИМА, с фокус върху вътрешната обработка на единиците, извлечени чрез частичен синтактичен анализ на повърхнинни езикови конструкции. В статия № 7 е показано как да се използва знанието за предметната област при анализа на медицински текст: *(i)* при сегментацията на текста и определяне на локална група клаузи, отнасящи се към дадено понятие; така анализът на текста

може да се извършва над малък фрагмент от клаузи, фокусирани над дадено понятие; и *(ii)* при запълване на извлечените от текста стойности в полетата на шаблона за съхранение на извлечената информация (с евентуално разширение на шаблона). Предложена е евристична стратегия за определяне кои понятия, намерени в текста, могат да се разглеждат като свързани с понятие, зададено във вътрешния шаблон. Разгледан е проблемът с липсата на концептуални ресурси на български език и е предложено адаптирано решение на базата на налични ресурси на английски език. В статия № 10 са предложени алгоритми за нормиране на стойности на признаци преди запълването им във вътрешните шаблони на системата. Нормирането помага за класификация на разнообразните думи, описващи състоянието на пациентите; например за състояние на кожата са намерени 93 различни описания, които трябва да бъдат класифицирани в категории *добро*, *леко увредено*, *увредено* и *силно увредено*. Така идентифицирането на показатели за състоянието на пациента е непосредствено свързано с категоризиране на състоянието спрямо медицинското значение на намерените признаци. Предложени са и стратегии за извличане на концептуални структури с български онтологични етикети от UMLS (the Unified Medical Language System) и по-специално от метатезауруса на UMLS. Тези стратегии са развити в статия № 12, която представя прототип за извличане на семантични връзки между понятия чрез синтактичен анализ на английски текстови дефиниции в метатезауруса на UMLS. Реализираният прототип извлича релацията *isa* между понятие и негово надпонятие, а също така и релацията *affects* между заболяване и засегнат от него орган. Прототипът избира дефиниция, от която да се извлече релация (тъй като в UMLS се съдържат различни медицински ресурси) и извършва известна обработка на парафразите с цел подготовка на текстовата дефиниция за прилагането на Link Parser. Всъщност липсата на семантични ресурси, в които има експлицитно-зададена релация *affects*, е едно от главните препятствия пред частичния семантичен анализ на болничния запис. Това обстоятелство се усложнява и от големия брой на понятия в медицината – десетки хиляди заболявания, органи, хиляди лекарства и т.н. Статия № 13 представя идеята за автоматично сегментиране на отделни епизоди от развитието на заболяването на пациента, които се разпознават в анамнезата на болничния запис поради явно-указани темпорални маркери. На практика анамнезата представлява едно резюме на историята на пациента, написано от експерт, който документира най-важните моменти от развитието на диабета. Представени са първи резултати от обработката на темпоралните маркери и тяхното автоматично разпознаване в 1197 болнични записа на диабетици. Проста линейна прогресия при нареждането на темпоралните маркери позволява и наредба на епизодите, които след последващи умозаключения могат да се наредят в интервали от събития.

Статии № 11, 14 и 15, публикувани през последните 5 месеца, представят резултати на българския колектив в проекта PSIP (Patient Safety through Intelligent Procedures in Medication). Колектив от ИИКТ-БАН под ръководството на Г. Ангелова участва в PSIP със задача да извлече от текста на болнични записи информация за диагнозите на пациента, за приложеното лекарствено лечение и за стойностите от клинични

изследвания, които не са направени в болницата. Тази задача е актуална поради факта, че в България пациентът внася лично в болницата и приема редица лекарства, които не са изписани през болничната аптека. Също така някои клинични изследвания (предимно хормонални тестове) не се извършват в болнични условия, а във външни лаборатории. Поради това автоматичният анализ на текста е единственият начин да се намерят данни за диагнозите, състоянието и лечението на пациентите. Статия № 11 разглежда резултатите от работата на трите екстрактора, които извличат диагнози, лекарства и стойности от клинични тестове от 6200 болнични записа на български език. Също така се представя подходът за интеграция на извлечените от текста стойности с данните, налични за съответните пациенти в болничната информационна система. При наличие на дублирани стойности като по-достоверни се вземат тези от информационната система. Статия № 14 представя интеграцията в повече детайли, като се спира върху приетите по премълчаване конвенции за определяне на времето, към което се отнасят извлечените от текста стойности (например клиничните тестове, описани само в текста, се датират към ден 0 от хоспитализацията, а приеманите лекарства за хронични заболявания се разпределят за всеки ден от престоя в болницата). Показан е и интерфейсът, интегриращ он-лайн екстракция на лекарства, създаден с цел валидиране на подхода на проекта PSIP в България. При приемането на пациент в болницата и въвеждане на неговата анамнеза в болничната система, екстракторът на лекарства предлага автоматично-извлечен списък на приеманите в момента на хоспитализацията лекарства. Лекарите одобряват такъв продукт, понеже той им спестява необходимостта да внасят ръчно предписания за лекарствено лечение. Статия № 15 резюмира експериментите, направени с цел извличане на имена на лекарства от различни зони на епикризите, с оглед максимално точно определяне на "текущото лечение" към момента на хоспитализацията. Показана е статистика, че имена на лекарства се срещат във всички зони на епикризата (дори при диагнозите се описват заболявания от типа на "хипотиреоидизъм, индуциран от амиодарон"). От текстовете на болничните записи в учебния корпус са научени автоматично различни фрази, указващи лекарства, приемани по време на хоспитализцията. С най-голяма точност "текущото" лечение се разпознава при автоматичен анализ на зоната "Анамнеза", поради което именно то е реализирано при подготовката на експерименталната база данни за валидиране на подхода на PSIP в България. Точността на разпознаване на лекарствата, приемани по време на хоспитализацията, е над 90%, като има и свръх-генерация в около 6% от случаите. До известна степен високата точност се дължи на структурата на епикризата, която позволява да се разпознае анамнезата в текста. Тази статия получи Rolf Hansen prize for best paper на Европейския конгрес по медицинска информатика в Осло през август 2011 г.

Статиите, публикувани през последните 1-2 години, показват и насоката за непосредствената бъдеща работа на Г. Ангелова, а именно:

- прецизиране на подхода за автоматично извличане на различни единици от текста на болнични записи,

- прилагане на създадените прототипни екстрактори към големи ресурси от болнични записи и
- създаване на методи, които се базират на автоматично извлечените показатели и осигуряват по-нататъшната им обработка с цел получаване на полезни медицински системи: за търсене на подобие между историите на заболяванията на различни пациенти, за намиране на странични ефекти на лекарства и т.н.

8 септември 2011 г.