Events
Хроника

ТРЕТА МЕЖДУНАРОДНА НАУЧНА КОНФЕРЕНЦИЯ „КОМПЮТЪРНАТА ЛИНГВИСТИКА В БЪЛГАРИЯ“ – ТРАДИЦИИ И ПЕРСПЕКТИВИ

Светлозара Лесева, Ивелина Стоянова
Институт за български език „Проф. Л. Андрейчин“ – БАН

На 28 и 29 май 2018 година в Дома на Европа и Централното управление на Българската академия на науките в София се проведе третото издание на Международната научна конференция „Компютърната лингвистика в България“ (Computational Linguistics in Bulgaria – CLIB 2018).

Конференцията е форум, на който български учени, работещи в областта на компютърната лингвистика в България и в чужбина, имат възможност да обменят познание, опит, идеи и постижения помежду си и с изследователи от други държави, работещи върху компютърнолингвистична проблематика, свързана с български език и други славянски и балкански езици, както и с представители на индустрията в областта на информационните технологии.

Тазгодишното издание на Конференцията премина при голям успех – значително разширено като тематика, брой участници и представителство на научната общност в областта на компютърната лингвистика. Представени бяха разработки на изследователи от България, Русия, Сърбия, Румъния, Япония, Украйна, Норвегия, Италия, САЩ, Великобритания. Разнообразната и балансирана научна програма бе организирана в няколко основни части: (а) пленарна сесия, в рамките на която бяха изнесени 3 лекции от компютърни лингвисти, заемащи водещи в световен мащаб позиции и гостуващи в България специално за участието си в CLIB 2018; (б) основна конференция с постерна сесия; и (в) специална научна сесия, посветена на лексикално-семантичните мрежи (от типа на УърдНет).

Конференцията бе открита на 27 май 2018 с публична лекция на проф. Руслан Митков – един от първите български компютърни лингвисти, учен с ярък принос в областта на автоматичната обработка на естествения език и езиковите технологии, ръководител на Изследователската група по компютърна лингвистика в Университета в Улвърхамптън, Великобритания. Лекци-

ята на тема „Езиковата интелигентност: компютрите срещу хората“ (*Linguistic Intelligence: Computers vs. Humans*) беше изнесена пред учени, гости на Конференцията, учители и ученици, участвали в състезанието по компютърна лингвистика, провело се в същия ден.

Пленарната сесия на CLIB 2018 започна с лекция на проф. д-р Руслан Митков на тема „С малко помощ от страна на компютърната обработка на естествения език: езикови технологии с ефект върху обществото“ (*With a Little Help from NLP: My Language Technology Applications with Impact on Society*). В лекцията си проф. Митков се спря на три разработени от него оригинални метода, които лежат в основата на реални езикови технологии, насочени в помощ на обществото и фокусирани в областта на електронното обучение, писмения и устния превод и подпомагането на лица с езикови затруднения. В първата част на лекцията си проф. Митков представи иновативна разработка за автоматично генериране на тестове с множествени изборни отговори от електронни учебници, които се основават на техники като извличане на термини, семантичен анализ и преобразуване на изреченията. Благодарение на създаденото приложение изготвянето на тестове става многократно по-бързо, без да се прави компромис с качеството. Втората област, която докладът засегна, бе създаването на преводна памет от ново поколение в помощ на професионалистите, занимаващи се с писмени преводи, а в бъдеще – и на колегите им, работещи в сферата на устния превод. В третата част от лекцията си проф. Митков запозна аудиторията с оригинална методология и система, ориентирана към улесняване на четенето и разбирането на текстове от лица, страдащи от аутизъм.

Вторият пленарен доклад на тема „Обучение с невронни графи“ (*Neural Graph Learning*) бе изнесен от д-р Суджит Рави – старши изследовател и мениджър в „Гугъл“. Д-р Рави представи иновативни разработки на алгоритми за машинно самообучение, базирани на теорията на графите, които благодарение на изчислителната и изразителната си мощ и простота правят възможно представянето на различни видове отношения, откривани в обработваните масиви от данни, и кодирането на структурни специфики, необходими за формулирането и решаването на реални научни и приложни задачи. Представената система комбинира алгоритми за самообучение, базирани на графи, с невронни мрежи за дълбоко обучение и може да оперира с огромни по обем данни при ниска сложност на операциите. Тази теоретико-приложна рамка е използвана в имплементацията на някои от най-съвременните потребителски ориентирани интелигентни приложения на компанията в областта на генерирането на автоматични отговори на електронни съобщения (Smart Reply), автоматичното разпознаване на образи, обобщаването на съдържанието на видео записи и др.

Третият пленарен доклад бе изнесен от д-р Зорница Козарева – мениджър в „Гугъл“ и лауреат на Наградата „Джон Атанасов“ за 2016 година за високи постижения в областта на компютърната лингвистика. В лекцията си „Създаване на интерактивни асистенти чрез дълбоко учене“ (*Building Conversational Assistants Using Deep Learning*) д-р Козарева запозна аудиторията с работата на екипа си върху технологии за създаване на интелигентни интерактивни асистенти, подпомагащи общуването между човек и компютър. Акцент в лекцията бе разрешаването на предизвикателствата пред създаването на помощници, които да са в състояние да извършват интелигентни решения, съобразени с индивидуалните предпочитания и навици на хората, с цел подпомагане на ежедневието им чрез организиране на графици, пътувания и др. Представени бяха решения на задачи като извличане на същности, предсказване на намеренията и отговаряне на въпроси, базирани на алгоритми за дълбоко самообучение. В заключение бяха очертани насоки на работа, основаващи се на нови открития в областта на предсказването на намеренията на потребителите, свързани с дейности като пазаруване, развлечения и под.

На Конференцията бяха изнесени две специални лекции, които очертаха друго актуално направление в областта на компютърната лингвистика, а именно развитието на езиковите умения и четивните способности в ранна детска възраст. В лекцията си „Оценка на четивната ефективност в началното училище, базирана на компютърната обработка на естествения език“ (*NLP-based Assessment of Reading Efficiency in Early Grade Children*) проф. Вито Пирели от Института по компютърна лингвистика „А. Замполи“ в Рим представи приложението на методи от обработката на естествения език при оценяване на уменията за четене. Проф. Мила Димитрова-Вълчанова и проф. Валентин Вълчанов от Норвежкия научно-технологичен университет в Трондхайм дискутираха особеностите на метафоричния език в контекста на индивидуалното езиково развитие и „разбирането“ на човешкия език от компютрите (*Figurative Language Processing: A Developmental and NLP Perspective*).

В рамките на основната конференция бяха изнесени доклади в следните тематични направления: автоматично резюмиране на текстове на новинарски статии за български език (Никола Таушанов, Иван Койчев и Преслав Наков – *Abstractive Text Summarization with Application to Bulgarian News Articles*); формализиране на лексикалното значение с помощта на онтологии (Мария Гриц – *Towards Lexical Meaning Formal Representation by Virtue of the NL-DL Definition Transformation Method*); специфични аспекти на словообразуването (Джуня Морита – *Narrow Productivity, Competition, and Blocking in Word Formation*); специфики на графичното представяне на текстовете с оглед на тяхната автоматична обработка (Цветана Крстев,

Ранка Станкович и Душко Витас – *Knowledge and Rule-Based Diacritic Restoration in Serbian*; Антон Зиновиев – *Perfect Bulgarian Hyphenation, or How not to Stutter at End-of-line*). В научна сесия, посветена на корпусните изследвания, бяха изнесени: доклад върху създаването на корпус и система за откриване на непреки анафори (Ана Ройтберг, Денис Хачко – *Russian Bridging Anaphora Corpus*); корпусно изследване върху видовите и времевите характеристики на миналите деятелни причастия в български език (Екатерина Търпоманова – *Aspectual and Temporal Characteristics of the Past Active Participles in Bulgarian – a Corpus-based Study*); изследване върху асиметрията при лексикализиране на наименования за лица от женски пол в български и украински (Олена Сирук, Иван Держански – *Unmatched Feminitives in a Corpus of Bulgarian and Ukrainian Parallel Texts*); корпус с резюмета на текстове за български език (Виктория Петрова – *The Bulgarian Summaries Corpus*).

Седем от приетите доклади бяха представени под формата на постери в рамките на постер сесия, предшествана от кратки устни представяния. Докладите включиха разработки в областта на композиционната дистрибуционна семантика (Амир Бакаров – *The Effect of Unobserved Word-Context Co-occurrences on a Vector-Mixture Approach for Compositional Distributional Semantics*), автоматичното разпознаване на авторство в текстови съобщения (Бранислава Шандрих – *Fingerprints in SMS Messages: Automatic Recognition of a Short Message Sender Using Gradient Boosting*), визуализацията на транскрибирана реч и нормализиран текст (Марина Джонова, Хетил Ро Хауге и Йовка Тишева – *Parallel Web Display of Transcribed Spoken Bulgarian with its Normalised Version and an Indexed List of Lemmas*), създаването на паралелен корпус за целите на статистическия автоматичен превод на глаголните времена между български и английски (Тодор Лазаров – *Bulgarian-English Parallel Corpus for the Purposes of Creating Statistical Translation Model of the Verb Forms. General Conception, Structure, Resources and Annotation*), езиковото обучение (Георги Джумайов – *Integrating Crowdsourcing in Language Learning*) и обучението по лингвистика и компютърна лингвистика чрез специфични видове задачи за ученици, състезаващи се в тези дисциплини (Иван Держански и Милена Венева – *Linguistic Problems on Number Names*; Росица Декова и Аделина Радева – *Introducing Computational Linguistics and NLP to High School Students*).

В рамките на Конференцията беше включена и специална сесия, посветена на УърдНет и онтологиите. Сесията бе организирана в рамките на проекта „Семантична мрежа с широк спектър от семантични релации“, изпълняван от Секцията по компютърна лингвистика към Института за български език и финансиран от Фонда за научни изследвания по програма „Финансиране на фундаментални научни изследвания“ за 2016 година.

Целта на специалната сесия бе да се създаде форум за споделяне на изследвания в областта на лексикално-семантичните мрежи и онтологиите и взаимодействието и интегрирането между двата типа представяне на знанието в ресурси с различна насоченост. Опитът и резултатите, които бяха обменени, предложиха ценни насоки за бъдещето на изследванията в тази област и имат пряко отношение към осъществяването на следващия етап от проекта.

Докладът на Наталия Лукашевич и Борис Добров (*Ontologies for Natural Language Processing: the Case of Russian*) представи група езикови ресурси за руски език, RuThes, основаващи се на обединяването на УърдНет с тезауруси и формални онтологии, като данните са представени в единен формат. Получените ресурси се използват в областта на компютърната обработка на естествения език и извличането на информация. Едно от реалните приложения на ресурса е полуавтоматичното генериране на РуУърдНет.

Разработката, представена от Ранка Станкович, Миляна Младенович, Иван Обрадович, Марко Витас и Цветана Крстев (*Resource-based WordNet Augmentation and Enrichment*), демонстрира подход за обогатяване на Сръбския уърднет с помощта на сръбско-английски ресурси. Методът се базира на превод и корекция на дефинициите от Принстънския уърднет на сръбски и автоматичен подбор на кандидати за членове на синонимните множества от списъци с преводни еквиваленти, извлечени от двуезикови ресурси. Представена е оценка на резултатите, при които се взема предвид обемът от корекции, извършени от експерти върху автоматично създадения вариант.

Докладът на колегите от Румъния (Мария Митрофан, Вержиника Барбу Митителу, Григорина Митрофан – *A Pilot Study for Enriching the Romanian WordNet with Medical Terms*) представи пилотно проучване, ориентирано към обогатяването на Румънския уърднет със специализирана лексика, поконкретно медицинска терминология. Статията изследва интеграцията на познанието от медицинския тезаурус SNOMED CT в йерархичната релационна структура на УърдНет и представя проблемни случаи, свързани с различната организация на познанието в двата ресурса.

В доклада на тема *Classifying Verbs in WordNet by Harnessing Semantic Resources* (Светлозара Лесева, Ивелина Стоянова и Мария Тодорова) бе представена класификация на глаголите в УърдНет, създадена автоматично чрез обединяването на преимуществата на три семантични ресурса – самия УърдНет и неговата разклонена йерархична структура, богатото и гранулирано семантично описание и таксономичните отношения във ФреймНет и пообобщеното семантично и синтактично базирано описание във ВърбНет. Въз основа на съотнасянето между трите ресурса и на вътрешната им структура се извлича класификация, чиито класификационни категории (семантичните класове) са пренесени от фреймовете (концептуалните структури) във ФреймНет, структурирани съобразно йерархичните отношения (хиперонимия/хипо-

нимия) в УърдНет. В резултат от създадената класификация са приписани автоматично и впоследствие ранкирани по вероятност класове на по-голямата част от синонимните множества в УърдНет.

Докладът на Ивелина Стоянова на тема *Factors and Features Determining the Inheritance of Semantic Primes between Verbs and Nouns within WordNet* изследва механизмите за наследяване на семантични свойства между деривационно свързани глаголи и съществителни и определя три типа наследяване между семантичните примитиви на глаголите и съществителните: универсални – независими от аргументната структура на глагола, които могат да са събитийни и обстоятелствени; общи – характерни за цели класове глаголи (напр. агентивни/неагентивни); специфични за конкретни глаголи – зависят от аргументната структура (както е представена в ресурси като ВърбНет и ФреймНет). В разработката са предложени възможности за разширяване на покритието на семантичните релации въз основа на информация за аргументната структура и се дискутират закономерностите при наследяването на семантични характеристики от глаголите към съществителните и прилагането им за разширяване на УърдНет със семантични множества, за формулиране на различни проверки на последователността на данните и мн.др.

В рамките на сесията бе представена и демонстрация на уеб базираната система за редактиране и визуализация на уърднети *Хидра* (Борислав Ризов и Цветана Димитрова – *Online Editor for WordNets*). Функционалностите на системата позволяват редактиране на синонимни множества в произволен брой уърднети, включително чрез добавяне или отстраняване на синоними, съставяне и редактиране на тълковни дефиниции, примери и друга информация, добавяне или отстраняване на релации и др.

Високата научна стойност на докладите, одобрени за представяне в рамките на CLIB 2018, бе гарантирана чрез процедурата за подбор чрез двойно анонимно рецензиране. Всяка статия бе оценена от трима независими рецензенти, изтъкнати специалисти в съответната научна област. В продължение на традицията сборникът с доклади от Международната научна конференция „Компютърната лингвистика в България“ да се включва в престижни бази от данни с научни публикации, сборникът от третото издание е предложен за индексиране в ISI Web of Knowledge. Сборникът с доклади, програмата и снимковият материал са достъпни и електронно на страницата на Конференцията (http://dcl.bas.bg/clib/).

За отразяването на събитието голяма роля изигра медийният партньор на Конференцията – Националното издателство за образование и наука „Аз-буки“. Сътрудничеството на Института за български език с различни представители на бизнеса в сферата на информационните технологии е предпоставка за бъдещото развитие и създаването на научни разработки в областта на компю-

търната лингвистика и иновативни и обществено полезни езикови технологии. Конференцията си поставя амбициозната задача и в бъдеще да съдейства за изграждане на мрежа за сътрудничество между български учени в страната и чужбина, работещи в областта на компютърната лингвистика, и учени, разработващи езикови технологии, приложими за българския език. Следващото издание на международната научна конференция „Компютърната лингвистика в България“ ще се проведе през 2020 година.

THIRD INTERNATIONAL SCIENTIFIC CONFERENCE “COMPUTER LINGUISTICS IN BULGARIA” – TRADITIONS AND PERSPECTIVES

✉ **Dr. Svetlozara Leseva, Dr. Ivelina Stoyanova**
Computer Linguistics Section
Institute for Bulgarian Language
Bulgarian Academy of Sciences
Sofia, Bulgaria
E-mail: zarka@dcl.bas.bg; iva@dcl.bas.bg