Резюмета на статиите за конкурса за професор
на
Кирил Симов

Доказателствен материал по група показатели В (научна област 4)

1. Staykova, K., Osenova, P., Simov, K.. New Applications of "Ontology-to-Text Relation" Strategy for Bulgarian. Cybernetics and Information Technologies, 12, 4, 2012, ISSN:1314-4081, SJR: 0.215

   The paper presents new applications of the Ontology-to-Text Relation Strategy to Bulgarian Iconographic Domain. First the strategy itself is discussed within the triple ontology-terminological lexicon-annotation grammars, then – the related works. Also, the specifics of the semantic annotation and evaluation over iconographic data are presented. A family of domain ontologies over the iconographic domain are created and used. The evaluation against a gold standard shows that this strategy is good enough for more precise, but shallow results, and can be supported further by deep parsing techniques.

   Статията представя ново приложение на подхода на релацията онтология-текст за тематичната област на българската иконография. Самата релация онтология-текст включва три компонента: онтология – речник - анотационна граматика. Представени са: а) особеностите на семантичното анотиране и б) оценката над текстове от тематичната област на иконографиката. Създадено е множество от онтологии за иконографиката. Оценката по отношение на корпус-еталон показва, че подходът е достатъчно добър за точни, но частични резултати. За да се подобри нивото на анотация, е необходимо да се използват дълбоки граматики.

2. Lemnitzer, L., Vertan, C., Killing, A., Simov, K., Evans, D., Cristea, D., Monachesi, P.. Improving the search for learning objects with keywords and ontologies. EC-TEL 2007: Creating New Learning Experiences on a Global Scale, LNCS, volume 4753, Springer, 2007, ISBN: 978-3-540-75194-6, DOI: https://doi.org/10.1007/978-3-540-75195-3_15, 202-216, SJR 0.283

   The paper reports on a current project results which aim at improving the effectiveness of retrieval and accessibility of learning object within learning management systems and learning object repositories. This task is approached by providing Language Technology based functionalities and by integrating semantic knowledge through domain-specific ontologies. The paper reports about the development of a keyword extractor and a domain-specific ontology, the integration of these modules into the learning management system ILIAS and the validation of these tools which assesses their added value in the scenario of searching learning objects across different languages.

   Статията представя резултатите на проекта LT4eL, които имат за цел да подобрят извличането на учебно съдържание и достъпа до него в една конкретна система за управление на обучението и хранилище на учебно съдържание. Този проблем е решен чрез предоставяне на достъп до функционалности, базирани на езикови технологии и чрез интегриране на семантично знание под формата на тематични онтологии. За целта са разработени: специализарана тематична онтология и екстрактор на ключови думи. Тези модули са интегрирани в системата за управление на обучението ILIAS. Беше направена оценка на тези добавени функционалности. Тя показа, че модулите подобряват резултатите в сценарий за търсене на учебно съдържание между няколко езика.

3. Vertan, C., Monachesi, P., Simov, K., Osenova, P., Lemnitzer, L., Killing, A., Evans, D.. Crosslingual retrieval in an eLearning environment. AI*IA 2007: AI*IA 2007: Artificial Intelligence and Human-Oriented Computing, LNCS, 4733, Springer, 2007, ISBN: Print ISBN 978-3-540-74781-9, DOI: https://doi.org/10.1007/978-3-540-74782-6_76, 839-847, SJR 0.283

   This paper reports on the project LT4eL (Language Technology for eLearning) aiming at improving the effectiveness of retrieval and accessibility of learning objects within a learning management system. It elaborates on the process of building the domain ontology and presents the multilingual support offered to the application. The ontology construction consists of the following steps: (1) Processing of the Keywords; (2) Formalization of the senses; and (3) Linking to upper ontologies. The result is a domain ontology linked to WordNet and DOLCE. Additionally lexicons in several languages are aligned to the ontology. For the annotation of the learning objects special annotation grammars were designed. The result was integrated within a learning management system to support the semantic search.

   Статията описва целите на проекта LT4eL за достъп до учебно съдържание в рамките на система за управление на обучението. Текстът представя процеса на създаване на тематична онтология и осигуряването на многоезичен достъп до нея, както и до учебния материал. Всички те се интегрират в системата. Изграждането на онтологията минава през следните етапи: (1) обработка на ключови думи и фрази; (2) формализиране на значенията в OWL; и (3) свързване на тематичната онтология към общи онтологии (upper ontology). Ключовите думи са извадени от многоезиково учебно съдържание, като се превеждат на английски. След това се оформят различните значения и техните дефиниции, които се формализират като понятия и релации, кодирани в OWL. Резултатът се съпоставя с WordNet и общата онтология DOLCE. Допълнително към онтологията се съпоставят речници на няколко езика за осигуряване на многоезиков достъп. За да се анотира учебното съдържание с понятия, от онтологията се изграждат специални анотационни граматики. Крайният резултат се интегрира към система за управление на обучението, за да поддържа възможностите за семантично търсене.

4. Osenova, P., Simov, K., Mossel, E.. Language Resources for Semantic Document Annotation and Crosslingual Retrieval. Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008, ELRA, 2008, 1873-1879

   This paper describes the interaction among language resources for an adequate concept annotation of domain texts in several languages. The architecture includes domain ontology, domain texts, language specific lexicons, regular grammars and disambiguation rules. This is considered the preparatory phase for the integration of a semantic search facility in Learning Management Systems. The implementation and performance of this search are discussed in the context of related work as well as other types of searches. Also the results from some preliminary steps towards evaluation of the concept-based and text-based search are presented.

   Статията описва взаимодействието на различни езикови ресурси, необходими за адекватното анотиране на текстове на няколко езика с тематични понятия. Архитектурата включва тематична онтология, тематични текстове, речници на различни езици, регулярни граматики и правила за снемане на многозначността. Тази архитектура се разглежда като предварителна за интегриране в система за управление на обучението. Имплементацията и качеството на търсене се дискутира по отношение на други типове търсене. Направено е сравнение между търсенето с понятия в анотирани текстове и пълнотекстовото търсене.

5. Lemnitzer, L., Simov, K., Osenova, P., Mossel, E., Monachesi, P.. Using a domain-ontology and semantic search in an eLearning environment. Innovative Techniques in Instruction Technology, E-Learning, E-Assessment, and Education, 2008, 279-284

The paper discusses the results of the integration of a semantic search module in a Learning Management System within "Language Technology for eLearning" (LT4EL) project. The project integrates the semantic knowledge in a Learning Management System to enhance the management, distribution and especially the cross-lingual retrieval of learning materials. One of the results achieved in the project is the construction of a language-independent domain-ontology with lexicons of eight languages linked to it. Learning objects of these languages have been annotated with concepts from this ontology. The ontology management system allows for semantic search which has been proven to be more effective than the simple full text search for all languages of the project.

Статията дискутира резултатите от интеграцията на модул за семантично търсене в система за управление на обучението в рамките на проекта "Language Technology for eLearning" (LT4EL). Модулът е насочен към анотирането на учебен материал и междуезиково търсене на учебни обекти. Един от резултатите на проекта е създаването на езиково независима онтология с подравнени към нея речници на осем езика. Учебен материал на всички тези езици е анотиран с понятия от онтологията. Системата за работа с онтологията позволява семантично търсене. Показано е, че това търсене е по-ефективно от обикновеното пълнотекстово търсене за всички езици в проекта.

6. Monachesi, P., Markus, T., Posea, V., Trausan-Matu, S., Osenova, P., Simov K.. Supporting knowledge discovery in an e-learning environment having social components.. Technological Developments in Networking, Education and Automation., 2010, 157-162

   One of the goals of the "Language Technology for LifeLong Learning" project is the creation of an appropriate methodology to support both formal and informal learning. Services are being developed that are based on the interaction between a formal representation of (domain) knowledge in the form of an ontology created by experts and a social component which complements it, that is tags and social networks. It is expected that this combination will improve learner interaction, knowledge discovery as well as knowledge co-construction.

   Статията представя резултати от проекта "Language Technology for LifeLong Learning", който има за една от своите основни цели да създаде подходяща методология да подкрепя формалното и неформалното образование. Разработени са услуги, които се основават на взаимодействието между формално представеното знание, каквато е домейн онтологията, и социалния компонент, който я допълва под формата на множество от етикети и достъп до социални мрежи. Очакванията са, че такава комбинация ще подобри общуването между обучаемите, откриването на ново знание, както и съвместното изграждане на знание.

7. Simov, K., Osenova, P.. Constructing of an Ontology-based Lexicon for Bulgarian. Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010, 2010, 3840-3844

   This paper reports on the progress in the creation of an Ontology-based lexicon for Bulgarian. The construction has started with the concept set from an upper ontology (DOLCE). Then it was extended with concepts selected from the OntoWordNet, which correspond to Core WordNet and EuroWordNet Base concepts. The underlying idea behind the ontology-based lexicon is its organization via two semantic relations - equivalence and subsumption. These relations reflect the distribution of lexical unit senses with respect to the concepts in the ontology. The lexical unit candidates for concept mapping have been selected from two large and well-developed lexical resources for Bulgarian - a machine readable explanatory dictionary and a morphological lexicon. In the initial step, the lexical units were handled that have equivalent senses to the concepts in the ontology (2500 at the moment). Then, in the second stage, lexical units were selected on their frequency distribution in a large Bulgarian corpus. This step was the more challenging one, since

it might require also additions of concepts to the ontology. This work was the beginning of the creation of BTB WordNet – BTB-WN.

Статията описва напредъка по създаването на български речник, основан на онтологии. Създаването му стартира с понятията от общата онтология DOLCE. Тя беше разширена с понятия, избрани от OntoWordNet, които съответстват на значенията в Core WordNet и множеството от основни понятия в EuroWordNet. Основната идея зад онтологично организирания речник е, че неговата организация е чрез две семантични релации – еквивалентност и включване. Тези релации отразяват разпределението на значенията на лексикалните елементи по отношение на понятията в онтологията. Лексикалните елементи за съпоставянето към понятия са избрани от два големи и добре изградени лексикални ресурса за българския – машинна версия на тълковен речник на българския и морфологичен речник. В първата стъпка на изграждане на ресурса бяха обработени около 2500 значения, които са еквивалентни на понятията в онтологията. Във втората стъпка избрахме думи според тяхната честота в голям корпус с текстове на български. Тази стъпка се оказа по-сложна, т.к. може да изисква добавянето на нови понятия в онтологията. Разработката е началото на изграждането на българския WordNet BTB – BTB-WN.

8. Osenova, P., Laskova, L., Simov, K. Exploring Co-reference chains for concept annotation of domain texts. Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010, 2010, 172-176

   The paper explores the co-reference chains as a way for improving the density of concept annotation over domain texts. The challenge is to enhance relations among concepts instead of text entities, the latter pursued in most works. Our ultimate goal is to exploit these additional chains for concept disambiguation as well as sparseness resolution at concept level. First, a gold standard was prepared with manually connected links among concepts, anaphoric pronouns and contextual equivalents. This step was necessary not only for test purposes, but also for better orientation in the co-referent types and distribution. Then, two automatic systems were tested on the gold standard. Note that these systems were not designed specially for concept chaining. The conclusion is that the state-of-the-art co-reference resolution systems might address the concept sparseness problem, but not so much the concept disambiguation task. For the latter, word-sense disambiguation systems have to be integrated.

   Статията представя работа, изследваща кореферентните вериги като механизъм за подобряване на гъстотата на анотиране с понятия на документи в определена тематична област. Задачата е да се разшири релацията между понятия вместо между текстови сегменти, както се прави в повечето публикации, посветени на тази тематика. Целта ни е да използваме тези вериги за снемане на многозначността на понятията, споменати в текста, и подобряване на гъстотата на анотиране с понятия. Първо, ръчно беше изграден корпус-еталон, където бяха добавени връзки между понятия, анафорични местоимения и контекстни еквивалетни изрази. Тази стъпка беше необходима не само за целите на тестването, но също и за по-доброто разбиране на типовете кореференция и техната дистрибуция. След това две автоматични системи бяха тествани над корпуса еталон. Тези две системи не бяха създадени специално за построяване на вериги от понятия. Заключенията от тези експерименти са, че най-добрите системи за кореферентни вериги са добри за разпространение на понятийната информация над текста, но не са достатъно добри за снемане на многозначността. За последната задача има нужда от разработването на специални системи за анотиране със значения.

Доказателствен материал по група показатели Г (научна област 4)

1. Damova, M., Kiryakov, A., Simov, K., Petrov, S. Mapping the central LOD ontologies to PROTON upper-level ontology. CEUR Workshop Proceedings. Volume 689, 2010. 5th International Workshop on Ontology Matching, OM-2010 - Collocated with the 9th International Semantic Web Conference, ISWC, Pages 61-72. SJR 0.166

   Linking Open Data (LOD) facilitates the emergence of a web of linked data by publishing and interlinking open data on the web in RDF. One can explore linked data across servers by following the links in the graph. This paper describes an approach to access these data by means of a single ontology, matched to the schemata describing several of the most common LOD datasets. They are presented in a reason-able view - FactForge (http://factforge.net) - the biggest and most heterogeneous body of factual knowledge on which inference is performed. Techniques of (a) making matching rules with "ontology expressions", (b) adding new instances with inference rules, and (c) extending the upper level ontology with classes and properties are employed. They succeed to align ontologies designed according to different principles and displaying conceptual and structural mismatches.

   Свързаните отворени данни (LOD) ускоряват появата на мрежа от свързани данни, публикувани и взаимосвързани чрез RDF. Тези данни могат да бъдат изследвани в рамките на различни сървъри, като се проследят връзките в построения граф. Статията описва един подход към достъпа до тези данни с помощта на една обща онтология, покриваща схемите, които описват няколко от най-популярни множества от данни в LOD. Тези множества от данни са показани в обединено представяне, позволяващо разсъждения над данните (reason-able view). Това общо представяне е качено на сайта FactForge (http://factforge.net). То представлява най-голямото и разнообразно множество от фактологично знание, което позволява извършването на изводи. Използват се следните техники: (а) имплементация на съпоставящи правила с изрази над онтологии; (б) добавянето на примери (инстанси) с правила за извод; и (в) разширяване на общата онтология с класове и свойства. Това позволява да се съпоставят онтологиите за различните множества, които са построени по различни правила и принципи, както и да се открият понятийни и структурни несъответствия.

2. Georgiev, G., Nakov, P., Ganchev, K., Osenova, P., Simov, K.. Feature-Rich Named Entity Recognition for Bulgarian Using Conditional Random Fields. International Conference Recent Advances in Natural Language Processing, RANLP, 2009, 113-117. SJR 0.143

   The paper presents a feature-rich approach to the automatic recognition and categorization of named entities (persons, organizations, locations, and miscellaneous) in news texts for Bulgarian. We combine well-established features used for other languages with language-specific lexical, syntactic and morphological information. In particular, we make use of the rich tagset annotation of the BulTreeBank (680 morpho-syntactic tags), from which we derive suitable task-specific tagsets (local and nonlocal). We further add domain-specific gazetteers and additional unlabeled data, achieving F1=89.4%, which is comparable to the state-of-the-art results for English.

   Статията описва подход, базиран на богато множество от характеристики, за автоматично разпознаване и категоризиране на наименовани същности (лица, организации, местоположения и други) в медийни текстове на български език. Комбинират се характеристики, използвани за други езици с езиково специфична лексикална, синтактична и морфологична информация. В частност се използва богатото анотиране на синтактичната база Бултрибанк (680 тага), от която изличаме подходящи специфични за задачата характеристики на локално и глобално ниво. Също така се добавят специфични списъци с имена за определени тематични области, както и допълнителна неанотирана информация. По този начин се постига резултат F1=89.4%, който е съпоставим с най-добрия резултат за английския език.

3. Simov, K., Osenova, P.. Towards Minimal Recursion Semantics over Bulgarian Dependency Parsing. Proceedings of the International Conference Recent Advances in Natural Language Processing 2011, 2011, ISSN:1313-8502, 471-478. SJR 0.143

   The paper discusses the transferring rules of the output from a dependency parser for Bulgarian into RMRS analyses. This task is required by the machine translation compatibility between Bulgarian and English resources. Since the Bulgarian HPSG grammar is still being developed, a repairing mechanism has been envisaged by parsing the Bulgarian data with the Malt Dependency Parser, and then retrieving RMRS analyses by exploring the linguistic knowledge within BulTreeBank-DP.

   Статията описва правила за преобразуване на резултата от дependентен парсер за българския език към т.нар. Стабилна минимална рекурсивна семантика - RMRS (Robust Minimal Recursive Semantics – семантично представяне на значението на изречения). Тази задача е необходима при уеднаквяването на езиковите ресурси за машинния превод между български и английски. Обикновено такива структури се строят от варианти на Опорната фразова граматика (ОФГ) (англ. HPSG) за дадените езици, но такава граматика за българския е все още в ранен стадий на развитие. Затова е необходимо да се предвиди механизъм за извличане на такива структури от текстове, парсирани с дependентен парсер (в случая Malt Dependency Parser). Правилата използват граматичните характеристики и синтактичните структури в базата със синтактични описания Бултрибанк-DP за построяването на анализите в RMRS структурите.

4. Osenova, P., Simov, K.. The Political Speech Corpus of Bulgarian. LREC 2012, ISSN: 978-2-9517408-7-7, 1744-1747

   The paper introduces the Political Speech Corpus of Bulgarian. First, its current state has been discussed with respect to its size, coverage, genre specification and related online services. Then, the focus goes to the annotation details. On the one hand, the layers of linguistic annotation are presented. On the other hand, the compatibility with CLARIN technical Infrastructure is explained. Also, some user-based scenarios are mentioned to demonstrate the corpus services and applicability.

   Статията описва корпуса с политическа реч на българския език. Дискутират се: неговата големина, покритие, особености на жанра и свързаните с него услуги. Фокусът е върху подхода към анотирането му. От една страна, са представените нивата на лингвистично моделиране. От друга страна, е описана съвместимостта с техническата инфраструктура CLARIN. Представени са примерни сценарии, които показват полезността му и свързаните с него услуги.

5. Chanev, A., Simov, K., Osenova, P., Marinov, S.. The bultreebank: Parsing and conversion. Recent Advances in Natural Language Processing V: Selected papers from RANLP, 309, 2007, 321-330, SJR 0.143

   Treebanks are often based on either of two grammatical formalisms: phrase structure (constituency) grammar or dependency grammar. However, sometimes it is necessary to transform treebank representations in order to test statistical parsers based on the alternative approach. This paper presents new parsing results for Bulgarian by training two statistical parsers (constituency and dependency) on BulTreeBank. In the paper the interaction between constituency and dependency representations in both the constituency and the dependency parser is explored using information based on the alternative formalism. It is shown that this interaction has a positive impact on parsing accuracy. The paper also presents an investigation of the relation between BulTreeBank and one of its dependency variants which had been automatically derived from the original treebank.

Синтактичните бази данни (treebanks) най-често използват един от двата граматични формализъма: фразовите (конституентни) граматики и граматиките на зависимостите (дependентни граматики). Понякога, за да се тестват синтактични анализатори (парсери) върху алтернативния подход, е необходимо да се трансформират синтактичните описания в рамките на този подход. Тази статия представя нови резултати за българския език, като се тренират два такива парсера (конституентен и дependентен) над синтактичната база Булtрибанк. Тук се изследва взаимодействието между конституентното и дependентното представяне и в двата вида парсери, като се използва информация от двата формализъма. Показва се, че това взаимодействие има положително влияние върху точността на парсирането. Статията също така показва и проучване на отношението между оригиналното представяне в Булtрибанк и един от нейните дependентни представяния.

6. Osenova, P., Simov, K., Laskova, L., Kancheva, S.. A Treebank-driven Creation of an OntoValence Verb lexicon for Bulgarian. Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), ELRA, 2012, ISBN:978-2-9517408-7-7, 2636-2640

   The paper presents a treebank-driven approach to the construction of a Bulgarian valence lexicon with ontological restrictions over the inner participants of the event. First, the underlying ideas behind the Bulgarian Ontology-based lexicon are outlined. Then, the extraction and manipulation of the valence frames is discussed with respect to the BulTreeBank annotation scheme and DOLCE ontology. Also, the most frequent types of syntactic frames are specified as well as the most frequent types of ontological restrictions over the verb arguments. The envisaged application of such a lexicon would be: in assigning ontological labels to syntactically parsed corpora, and expanding the lexicon and lexical information in the Bulgarian Resource Grammar.

   Статията описва подход за построяване на български валентен речник с онтологични ограничения върху участниците в обозначеното събитие на базата на синтактичната база данни Булtрибанк. Първо, в статията са описани идеите, които са в основата на българския онтологично базиран речник. След това са представени извличането и обработката на валентните рамки в съответствие с анотирането в рамките на Булtрибанк и общата онтология DOLCE. След извличането на синтактичните дървета за дадена лема, се минава през обобщен синтактичен анализ, който отразява структурата на събитието на това ниво. Следващата стъпка е заместването на думите с понятия. Дискутирани са и приложенията на речника, които са свързани най-вече с подпомагане на синтактичния анализ и разширяването на други ресурси.

7. Savkov, A., Laskova, L., Kancheva, S., Osenova, P., Simov, K.. Linguistic Analysis Processing Pipeline for Bulgarian. Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), ELRA, 2012, 2959-2964

   This paper presents a linguistic processing pipeline for Bulgarian including morphological analysis, lemmatization and syntactic analysis of Bulgarian texts. The morphological analysis is performed by three modules — two statistical-based and one rule-based. The combination of these modules achieves the best result for morphological tagging of Bulgarian over a rich tagset (680 tags). The lemmatization is based on rules, generated from a large morphological lexicon of Bulgarian. The syntactic analysis is implemented via MaltParser. The two statistical morphological taggers and MaltParser are trained on datasets constructed within BulTreeBank project. The processing pipeline includes also a sentence splitter and a tokenizer. All tools in the pipeline are packed in modules that can also perform separately. The whole pipeline is designed to be able to serve as a back-end of a web service oriented interface, but it also supports the user tasks with a command-line interface. The processing pipeline is compatible with the Text Corpus Format, which allows it to delegate the management of the components to the WebLicht platform.

Статията представя последователността от обработващи модули за българския език, които извършват сегментиране на текста, морфологичен анализ, лематизация и синтактичен анализ. Морфологичният анализ е извършен с помощта на три модула: два статистически тагера и един модул, базиран на правила. Комбинацията от тези модули демонстрира най-добър резултат над богато множество от тагове (680). Лематизацията е имплементирана с помощта на голям морфологичен речник. Синтактичният анализ е имплементиран с помощта на MaltParser. Статистическите морфологични модули и MaltParser са тренирани над данните от Бултрибанк. Модулите могат да се използват както заедно, така и поотделно, което позволява интегрирането им с други модули. Цялостният модул може да се използва чрез уебуслуга, но може да се използва и като команда в операционната система. Използва се форматът Text Corpus Format (TCF), който е част от платформата WebLicht, разработена в рамките на немския CLARIN.

8. Georgiev, G., Zhikov, V., Simov, K., Osenova, P., Nakov, P.. Feature-Rich Part-of-speech Tagging for Morphologically Complex Languages: Application to Bulgarian. Proceedings of EACL 2012 - 13th Conference of the European Chapter of the Association for Computational Linguistics, ACL, 2012, 492-502

   The paper presents experiments with part-of-speech tagging for Bulgarian, a Slavic language with rich inflectional and derivational morphology. Unlike most previous work, which has used a small number of grammatical categories, we work with 680 morpho-syntactic tags. We combine a large morphological lexicon with prior linguistic knowledge and guided learning from a POS-annotated corpus, achieving accuracy of 97.98%, which is a significant improvement over the state-of-the-art for Bulgarian.

   Статията описва експерименти за морфологично анотиране за българския език, който е морфологично богат език по подобие на другите славянски езици. За разлика от предишни разработки, тук се работи с пълното множество от 680 морфологични тага. Използва се голям морфологичен речник с кодирано лингвистично знание и модул за насочено машинно сомообучение, който се стреми да реши първо по-лесните случаи на многозначност. Обучението се извършва над ръчно анотиран корпус. Постигната точност е 97.98%, което е значително подобрение на резултата за българския език.

9. Simov, K., Osenova, P., Georgiev, G., Zhikov, V., Tolosi, L. 2012. Bulgarian Question Answering for Machine Reading. Question Answering for Machine Reading Evaluation (QA4MRE), part of CLEF 2012. CEUR Workshop Proceedings, Volume 1178, ISBN:978-88-904810-3-1. SJR 0.166

   In CLEF 2012 the BulTreeBank Group of LMD, IICT, BAS participated in the QA4MRE (Question Answering for Machine Reading Evaluation) task for Bulgarian. The system represented in the paper exploits an NLP Pipeline for Bulgarian in order to process the questions, answers and the supporting texts. Then we represent the results of the analysis as a bag of linguistic units - lemmas, dependency relations. These bags of words are the match between the question plus answer and the sentences in the text. The answer that maximizes the overlap is selected as the correct one. Since the system is deterministic we have only one run. The score achieved by the run is 0.29. The other two runs are performed as baseline runs with randomly selected answers. Their scores are 0.20 and 0.12, respectively. Thus, the using of linguistic units in the overlapping estimation provides significant improvements over the baseline.

   CLEF е инициатива за изграждане на данни и провеждане на общи задачи в областта на отговорите на въпроси и изличането на информация над различни по тип данни – текстови, мултимедийни и други. Статията е свързана с участието ни в задачата QA4MRE (Question Answering for Machine Reading Evaluation – Отговаряне на въпроси за целите на оценяването на машинното четене) за български. Описаната система използва модул за обработка на български текстове. Този модул се използва за обработка на въпросите, отговорите и поддържащите текстове. След това резултатите от тези обратки се представят като ненарадена колекция от езикови елементи – леми и депендентни релации. Тези

колекции представляват съпоставянето между въпроси и отговори с изреченията от текста. Отговорът с максимално покритие се избира за коректен. Тъй като системата е детерминистична, то няма различни експерименти. Постигнатият резлтат е 0.29. Другите експерименти са като базов резултат със случайно избрани отговори. Техните стойности са съответно 0.20 и 0.12. Експериментът показва, че използването на лингвистична информация значително подобрява резултата.

10. Zhikov, V., Tolosi, L., Osenova, P., Simov, K., Georgiev, G.. Cross-Language Answer Validation. Question Answering for Machine Reading Evaluation (QA4MRE), part of CLEF 2012. CEUR Workshop Proceedings, Volume 1178, ISBN:978-88-904810-3-1. SJR 0.166

    The paper describes three language-independent methods for the task of answer validation. All methods are based on a scoring mechanism that rejects the degree of similarity between the question-answer pairs and the supporting text. The proposed methods are evaluated using various string similarity metrics, such as exact matching, Levenshtein, Jaro and Jaro-Winkler. In addition to this baseline approach, we take advantage of the multilingual QA4MRE (Question Answering for Machine Reading Evaluation) dataset, and devise an ensemble method, which chooses the answer indicated as correct by the largest number of analyses of the individual translations. Finally, we present a language-augmented method that enriches the questions and answers with paraphrases obtained by means of machine translation. Our methods depend on parameters which we estimate using the dataset from CLEF2011. The paper shows that all of the described approaches achieve a significant improvement over the random baseline, and that both majority voting and language augmentation lead to superior accuracy as compared with the original method.

    Статията представя три езиково независими метода за решаване на задачата за валидиране на отговори. И трите метода са базирани на ранкиращ механизъм, който отхвърля степента на подобие между двойките въпрос-отговор и поддържащите текстове. Предложените методи са оценени чрез използване на различни методи за подобие на низове – точно съвпадение, разстояние на Левенщайн, Яро и Яро-Винклер. В допълнение към тези базови модели се използват многоезковите данни на QA4MRE (Question Answering for Machine Reading Evaluation) за имплементация на асемблиращ метод, който избира отговора, маркиран като правилен чрез най-голямото множество от анализи на отделните преводи. Накрая е представен езиково обогатен метод, който към въпросите и отговорите добавя парафрази, получени чрез машинен превод. Тези методи зависят от параметри, които бяха надстроени на базата на данните от състезанието CLEF 2011. Резултатите показват, че всички подходи подобряват значително базовата система, получена чрез случаен избор. Също така се показва, че методът на мнозинството и добавянето на езикова информация постигат по-висока точност в сравнение с първоначалния подход.

11. Bishop, B., Kiryakov, A., Tashev, Z., Damova, M., Simov, K.. OWLIM Reasoning over FactForge. Proceedings of OWL Reasoner Evaluation Workshop (ORE'2012), collocated with IJCAR 2012, CEUR Workshop Proceedings, Vol-858, 2012, ISSN:1613-0073. SJR 0.166

    This paper presents the reasoning mechanism in the OWLIM family of semantic repositories, which is based on materialization – addition of all facts that follow from the current data and the implemented inference rules. This mechanism is evaluated using a combination of datasets from the Linked Open Data cloud in a public service called FactForge, where the benefits of materialization are manifested in improved SPARQL query performance.

    Статията разглежда механизмите за извод, имплементирани в едно семейство от семантични хранилища на RDF данни, които носят общото име OWLIM. Тези правила използват подхода на материлизицията – добавяне на всички факти, които следват от текущите данни и от имплементираните правила за извод. Този подход е оценен над комбинация от множества данни от колекцията от свързани отворени данни – Linked Open

Data Cloud. Тази комбинация е наречена FactForge (http://factforge.net/) и чрез нея се показва, че материлизацията подобрява изпълнението на SPARQL заявките.

12. Damova, M., Simov, K., Tashev, Z., Kiryakov, A. 2012. FactForge: Data Service or Diversity through Inferred Knowledge over LOD. In Proceedings of AIMSA'2012. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 7557 LNAI, pp. 145-151. ISSN: 03029743 ISBN: 978-364233184-8 DOI: 10.1007/978-3-642-33185-5_16, SJR 0.283

    Linked Open Data movement is maturing. LOD cloud increases by billions of triples yearly. Technologies and guidelines about how to produce LOD facts, how to assure their quality, and how to provide vertical oriented data services are being developed. Little is said however about how to include reasoning in the LOD framework, and about how to cope with its diversity. This paper deals with this topic. It presents a data service – FactForge – the biggest body of general knowledge from LOD on which inference is performed. It has close to 16B triples available for querying, derived from about 2B explicit triples, after inference and some OWLIM repository specific optimization. We discuss the impacts of the reference layer of FactForge and inference on the diversity of the web of data, and argue for a new paradigm of data services based on linked data verticals, and on inferred knowledge.

    Статията отчита напредъка в областта на свързаните отворени данни (Linked Open Data - LOD), където самата колекция от отворени данни нараства с милиарди от RDF факти във вид на тройки всяка година. Развиват се технологиите и насоките за създаване на LOD факти, както и за осигуряването на тяхното качество и предоставянето на услуги над тях. Но малко се дискутира въпросът за правене на извод на тези данни, които са разнообразни, имат различни концептуализации и са разположени на различни сървъри. Статията описва услугата за достъп към такива данни, наречена FactForge – най-голямото множество от общо знание от отворени свързани данни, над които са приложени правила за извод. FactForge (в този момент) се състои от 16 милиарда RDF тройки, над които може да се извършва извличане с помощта на SPARQL заявки. Тези 16 милиарда тройки са извлечени от два милиарда експлицитни тройки, след прилагане на правилата за извод и оптимизации, специфични за OWLIM хранилищата – например специфична обработка на транзитивните релации, където броят на тройките може да нарасне експоненциално. В статията се дискутира референтното ниво на FactForge и изводът, приложен към разнообразни данни от мрежата. Това референтно ниво се състои от онтологии, които допускат унифициран достъп до свързаните отворени данни.

13. Zhikov, V., Georgiev, G., Simov, K., Osenova, P. 2013. Combining POS tagging, dependency parsing and co-referential resolution for Bulgarian. International Conference Recent Advances in Natural Language Processing, RANLP. pp. 755-762. ISSN: 13138502, SJR 0.143

    This paper proposes a combined model for POS tagging, dependency parsing and co-reference resolution for Bulgarian — a pro-drop Slavic language with rich morphosyntax. We formulate an extension of the MSTParser algorithm that allows the simultaneous handling of the three tasks in a way that makes it possible for each task to benefit from the information available to the others, and conduct a set of experiments against a treebank of the Bulgarian language. The results indicate that the proposed joint model achieves state-of-the-art performance for тхе POS tagging task, and outperforms the current pipeline solution.

    В тази разработка се предлага обединен модел за едновременното решаване на три задачи при обработката на естествен език: граматичното тагиране, депендентното парсиране и изграждането на кореферентни вериги за българския – славянски език с богата морфология и нулева субектност. Формулирано е разширение на алгоритъма на парсиране с минимални покриващи дървета, което позволява решаването на трите задачи по такъв начин, че те да си взаимодействат в процеса на решение. Експериментите са извършени с данни от синтактическата база на българския език Бултрибанк. Резултатите показват, че моделът

постига най-добрите си резултати при: а) граматичното тагиране в частност и б) цялостния обработващ модул.

14. Ghayoomi, M., Simov, K., Osenova, P. 2014. Constituency parsing of Bulgarian: Word-vs. Class-based parsing. Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014. pp. 4056-4060.

    This paper reports the obtained results of two constituency parsers trained with BulTreeBank, an HPSG-based treebank for Bulgarian. To reduce the data sparsity problem, we propose using the Brown word clustering to do an off-line clustering and map the words in the treebank to create a class-based treebank. The observations show that when the classes outnumber the POS tags, the results are better. Since this approach adds on another dimension of abstraction (in comparison to the lemma), its coarse-grained representation can be used further for training statistical parsers.

    Статията описва резултатите от експериментите с два конституентни парсера, тренирани над базата със синтактични описания Бултрибанк. За да се намали влиянието на недостатъчните данни (законът на Зипф за разпределението на езиковите елементи в текстовете), се използва методът на Браун за клъстеризирането на думи, който съпоставя думите от базата със синтактични описания с класовете от клъстърите. По този начин се получават синтактични описания с класове. Експериментите показват, че когато броят на класовете е по-голям от броя на граматичните тагове в синтактичните описания, резултатите са по-добри. Според нас класовете допринасят за класификация на думите на друго (по-абстрактно) ниво в сравнение с лемите. Така това не много детайлно представяне има потенциал за подобряване на резултатите от такъв вид парсери.

15. Simov, K., Simova, I., Ivanova, G., Mateva, M., Osenova, P. 2014. A system for experiments with dependency parsers. Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014. pp. 4061-4065

    In this paper we present a system for experimenting with combinations of dependency parsers. The system supports initial training of different parsing models, creation of parsebank(s) with these models, and different strategies for the construction of ensemble models aimed at improving the output of the individual models by voting. The system employs two algorithms for construction of dependency trees from several parses of the same sentence and several ways for ranking of the arcs in the resulting trees. We have performed experiments with state-of-the-art dependency parsers including MaltParser, MSTParser, TurboParser, and MATEParser, on the data from the Bulgarian treebank – BulTreeBank. Our best result from these experiments is slightly better then the best result reported in the literature for this language.

    В тази статия се описва система за експериментиране с комбинации от депендентни парсери. Идеята е да се комбинират резултатите от различни модели. Системата поддържа следните стъпки: тренирането на различни модели; създаването на бази данни с автоматично парсирани изречения; различни стратегии за конструирането на асемблиращи модели с цел подобряване на резултата от различните модели чрез избор на решение в даден контекст. Системата поддържа два алгоритъма за конструиране на синтактичното дърво от различните анализи за едно и също изречение. Също така, системата предлага различни начини на ранкиране на конкуриращи се дъги в синтактичното дърво. Двата алгоритъма се различават по начина на оптимизация на ранкирането – единият е локален, а другият е глобален. Извършени са експерименти със следните популярни по онова време парсери: MaltParser, MSTParser, TurboParser и MATEParser. Техните модели са тренирани над данни от базата със синтактични описания – Бултрибанк. Най-добрият резултат от експериментите е малко по-добър от известните предишни резултати. Друго ценно наблюдение е, че асемблираният резултат не е монотонен по отношение на броя на първоначалните модели, което означава, че има нужда от изчерпване на различните комбинации. За целта системата беше пусната да работи в паралел.

16. Todorova, V., Simov, K. 2015. Training automatic transliteration models on DBpedia data. International Conference Recent Advances in Natural Language Processing, RANLP. 2015, pp. 654-662. ISSN: 13138502, SJR 0.143

    Our goal is to facilitate named entity recognition in Bulgarian texts by extending the coverage of DBpedia (http://www.dbpedia.org/) for Bulgarian. For this task we have trained translation Moses models to transliterate foreign names to Bulgarian. The training sets were obtained by extracting the names of all people, places and organizations from DBpedia and its extension Airpedia (http://www. airpedia.org/). Our approach is extendable to other languages with small DBpedia coverage.

    В задачата за разпознаване на наименовани същини (хора, организации, локации и други) обикновено се използват списъци с имена на познати такива същности. Един от източниците на такива имена е базата DBpedia (http://www.dbpedia.org/), построена на базата на Уикипедията за различни езици. За съжаление, DBpedia за български е сравнително малка. Нашата цел в тази разработка е да разширим нейното покритие с транскрибирани имена от DBpedia бази за други езици. Траслитерацията беше имплементирана като система за превод на имената по букви. Създадени бяха паралелни списъци с имена от DBpedia и Airpedia. Върху тях беше тренирана системата за машинен превод Moses. Този подход може да се приложи и за други езици, които нямата добро покритие на DBPedia.

17. Simov, K., Popov, A., Osenova, P. 2015. Improving word sense disambiguation with linguistic knowledge from a sense annotated treebank. International Conference Recent Advances in Natural Language Processing, RANLP 2015. pp. 596-603. ISSN: 13138502, SJR 0.143

    In this paper we present an approach for the enrichment of WSD (Word Sense Disambiguation) knowledge bases with data-driven relations from a gold standard corpus (annotated with word senses, valency information, syntactic analyses, etc.). We focus on Bulgarian as a use case, but our approach is scalable to other languages as well. For the purpose of exploring such methods, the Personalized Page Rank algorithm was used. The reported results show that the addition of new knowledge improves the accuracy of WSD with approximately 10.5%.

    Задачата за анотиране на думите в текста със значения - WSD (Word Sense Disambiguation) - стана особено популярна в последните десетина години благодарение на организираните състезания за нейното решаване. Единият от подходите е базиран на граф със знание, където чрез алгоритми за блуждаене над графа (random walk on graphs algorithms) се избират най-подходящите значения в контекста на текста. В статията се описва подход за обогатяване на графа със знания с нови дъги (наричани още релации), извлечени от корпуси-еталони, анотирани със значения, валентна информация и синтактични анализи. В тази статия фокусът е върху българския език, но този подход е приложим и към други езици. В експериментите беше показано, че добавянето на ново знание под формата на релации подобрява точността с 10.5 %.

18. Boella, G., Di Caro, L., Graziadei, M., Cupi, L., Salaroglio, C.E., Humphreys, L., Konstantinov, H., Marko, K., Robaldo, L., Ruffini, C., Simov, K., Violato, A., Stroetmann, V. 2015. Linking legal open data: Breaking the accessibility and language barrier in European legislation and case law. Proceedings of the International Conference on Artificial Intelligence and Law. 08-12-June-2015, pp. 171-175

    In this paper we describe how the EUCases FP7 project is addressing the problem of lifting Legal Open Data to Linked Open Data to develop new applications for the legal information provision market by enriching structurally the documents (first of all with navigable references among legal texts) and semantically (with concepts from ontologies and classification). First we describe the social and economic need for breaking the accessibility barrier in legal information in the EU, then we describe the technological challenges and finally we explain how the EUCases project is addressing them by a combination of Human Language Technologies.

В тази статия са описани резултати от европейския проект EUCases за обработка на правни документи с цел на създаване на множество със свързани отворени данни в правната тематична област. Целта е да се разработят нови приложения, използващи обогатени правни документи с понятия от онтологии, търсенето в тях и тяхното класифициране. В статията се дискутират проблемите на разпространение на правна информация на европейско ниво в контекста на технологичните предизвикателства. Представени са решенията, имплементирани в проекта. Използват се езикови технологии за анотиране с понятия от онтологии с правни понятия и имена (Geonames). След анотацията се използва модул за извличане на тези анотации във формата на RDF тройки, които се записват в множеството от свързани данни. Използването на готови онтологии и бази отворени данни осигурява свързването на новопостроената база със съществуващите такива.

19. Simov, K., Osenova, P., Popov, A. 2016. Using context information for knowledge-based word sense disambiguation. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 9883 LNAI, pp. 130-139. DOI: 10.1007/978-3-319-44748-313, SJR 0.283

    One of the most successful approaches to Word Sense Disambiguation (WSD) in the last decade has been the knowledge-based approach, which exploits lexical knowledge sources such as WordNets, ontologies, etc. The knowledge encoded in them is typically used as a sense inventory and as a relations bank. However, this type of information is rather sparse in terms of senses and the relations among them. In this paper we present a strategy for the enrichment of WSD knowledge bases with data-driven relations from a gold standard corpus (annotated with word senses, syntactic analyses, etc.). We focus on English as use case, but our approach is scalable to other languages. The results show that the addition of new knowledge improves the accuracy of WSD task with near 10%.

    Един от успешните модели за снемане на многозначността на думите в текстове (или анотациите със значения) – Word Sense Disambiguation (WSD) – е подходът, базиран на знание. Той използва лексикална база знание (например WordNet или лексикализирани онтологии). Знанието, кодирано в тези ресурси, се използва като източник на значения на думите и като релации между тези значения. Обикновено този тип информация е ограничен по отношение на значенията, но най-вече по отношение на релациите между тях. Обикновено те предоставят само дефиниционните характеристики на знанието, но не и неговото проявление в контекста на дадена ситуация например. В тази статия се представя подход към добавяне на контекстуална информация от корпуси-еталони с цел подбряване на базата със знание. Тук фокусът е върху английския език, т.к. той има необходимите ресурси, но подходът е приложим и за други езици. Експериментите показват подобрение на точността с близо 10 %.