РЕЗЮМЕТА

на научните публикации

на Стоян Милков Михов

за участие в конкурса за академичната длъжност „професор“

обявен в ДВ бр. 45/28.05.2021 г.

[1] Mihov, S., Schulz, K.U.
Efficient dictionary-based text rewriting using subsequential transducers
(2007) Natural Language Engineering, 13 (4), pp. 353-381.
ISSN: 13513249
ABSTRACT: Problems in the area of text and document processing can often be described as text rewriting tasks: given an input text, produce a new text by applying some fixed set of rewriting rules. In its simplest form, a rewriting rule is given by a pair of strings, representing a source string (the "original") and its substitute. By a rewriting dictionary, we mean a finite list of such pairs; dictionary-based text rewriting means to replace in an input text occurrences of originals by their substitutes. We present an efficient method for constructing, given a rewriting dictionary D, a subsequential transducer T that accepts any text t as input and outputs the intended rewriting result under the so-called "leftmost-longest match" replacement with skips, t'. The time needed to compute the transducer is linear in the size of the input dictionary. Given the transducer, any text t of length |t| is rewritten in a deterministic manner in time O(|t| + |t'|), where t' denotes the resulting output text. Hence the resulting rewriting mechanism is very efficient. As a second advantage, using standard tools, the transducer can be directly composed with other transducers to efficiently solve more complex rewriting tasks in a single processing step.
РЕЗЮМЕ: Проблемите в областта на обработката на текст и документи често могат да бъдат описани като задачи за пренаписване на текст: като се въведе входящ текст, се създава нов текст, като се прилага някакъв фиксиран набор от правила за пренаписване. В най-простата си форма правилото за пренаписване се дава от двойка низове, представляващи изходен низ („оригиналът“) и неговия заместител. Под пренаписване на речник имаме предвид краен списък на такива двойки; презаписване на текст, базиран на речник, означава да се заменят във входящ текст срещанията на изходните низове с техните заместители. В тази работа ние представяме ефективен метод, който по даден речник за презапис D конструира подпоследователен преобразувател T, който приема всеки текст t като вход и извежда желания резултат от презапис t' при така наречената стратегия замяна на „най-ляво най-дълго срещане“ с пропускания. Времето, необходимо за конструиране на преобразувателя, е линейно по размер на входния речник. Като се използва преобразувателят, всеки текст t с дължина |t| се презаписва по детерминиран начин за време O(|t| + |t'|), където t' е получения изходен текст. Следователно полученият механизъм за пренаписване е е с оптимална ефективност. Като второ предимство, използвайки стандартни инструменти, преобразувателят може да бъде директно композиран с други преобразуватели за ефективно решаване на по-сложни задачи за презаписване в една стъпка на обработка.

[2] Gerdjikov, S., Mihov, S., Schulz, K.U.
Space-efficient bimachine construction based on the equalizer accumulation principle
(2019) Theoretical Computer Science, 790, pp. 80-95.
ISSN: 03043975

ABSTRACT: Algorithms for building bimachines from functional transducers found in the literature are based on the following principle: each run of the bimachine simulates a particular successful path of the input transducer. Every single bimachine output exactly corresponds to the output of a single transducer transition. Here we introduce an alternative construction principle called the equalizer accumulation principle. It suggests that the bimachine steps take into account alternative parallel transducer paths, maximizing the possible output at each step using a joint view. This results in a construction where the deterministic left and right automaton of the bimachine both have size bounded by $2^{|Q|}$ where |Q| is the number of transducer states. In contrast, previous bimachine constructions lead to larger automata. We present a class of real-time functional transducers with n+2 states for which the standard bimachine construction generates a bimachine with at least Θ(n!) states whereas the construction based on the equalizer accumulation principle leads to 2^n+n+3 states. On the other end we present a real-time functional transducers with 4(n+1) states that cannot be represented as a bimachine with less than 2^n states. Therefore the space complexity of our construction is close to optimal in terms of the number of states. The new construction can be applied to rational functions from free monoids to “mge monoids”, a large class of monoids including free monoids, groups, and others that is closed under Cartesian products.
Резюме: Съществуващите в литературата алгоритми за строене на бимашини от функционални преобразуватели се основават на следния принцип: всяко изпълнение на бимашината симулира конкретен успешен път на изходния преобразувател. Всеки отделен (частичен) изход на бимашината съответства на конкретен преход на преобразувателя. В тази статия въвеждаме алтернативен принцип за строене на бимашини, който наричаме принцип за акумулиране на изравнители. Според него, отделните стъпки на бимашината отчитат различните възможни (успешни) пътища на преобразувателя, като максимизират възможния изход по всички тях. Това води до конструкция, при която и левият и десният автомат на бимашината имат размер, ограничен от $2^{|Q|}$, където |Q| е броят на състоянията в първоначалния преобразувател. За сравнение, предишните конструкции на бимашини водят до асимптотично по-големи автомати. В настоящата статия, показваме клас от реалновременни функционални преобразуватели с n+2 състояния, за който стандартната конструкция генерира бимашина с поне Θ(n!) състояния докато конструкцията, която следва принципа за акумулиране на изравнители, води до 2^n+n+3 състояния. От друга страна, представяме клас от реалновременни функционални преобразуватели с 4(n+1) състояния, който не допуска бимашина с по-малко от 2^n състояния. Това показва, че пространствената сложност на нашата конструкция се приближава до оптималната. Новопредложената конструкция може да се прилага за рационални функции с дефиниционна област в свободен моноид и област от стойности в "mge моноиди" – широк клас от моноиди, който включва свободните моноиди, групите и други, и който е затворен относно декартово произведение.

[3] Geneva, D., Shopov, G., Mihov, S.
Algorithms for Probabilistic and Stochastic Subsequential Failure Transducers
(2021) Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 12803 LNCS, pp. 127-139.
ISBN: 9783030791209
ABSTRACT: This paper introduces a framework for building probabilistic models with subsequential failure transducers. We first show how various types of subsequential transducers commonly used in natural language processing are represented by probabilistic and conditional probabilistic subsequential failure transducers. Afterwards we introduce efficient algorithms for composition of conditional probabilistic subsequential transducers with probabilistic subsequential failure transducers and weight pushing (canonization) of

probabilistic subsequential failure transducers. Those algorithms are applicable to many tasks for representing probabilistic models with subsequential failure transducers. One such task is the construction of the HCLG weighted transducer used in speech recognition which we describe in detail. At the end, empirical results and comparison between the presented HCLG failure weighted transducer and the standard HCLG weighted transducer constructions are shown.
РЕЗЮМЕ: Тази статия въвежда методология за конструиране на вероятностни модели с подпоследователни преобразуватели с преходи при неуспех. Първо показваме как различни типове подпоследователни преобразуватели, често използвани при обработката на естествен език се представят като вероятностни и условно-вероятностни подпоследователни преобразуватели с преходи при неуспех. След това въвеждаме ефективни алгоритми за композиция на условно-вероятностен подпоследователен преобразувател с вероятностен подпоследователен преобразувател с преходи при неуспех, както и алгоритми за изтласкване на теглата (канонизиране) на вероятностни преобразуватели с преходи при неуспех. Тези алгоритми са приложими за много задачи за представяне на вероятностни модели с подпоследователни преобразуватели с преходи при неуспех. Една такава задача е конструкцията на HCLG претегления преобразувател, използван при разпознаване на реч, който ние описваме подробно. В края са показани емпирични резултати и сравнение между представения HCLG претеглен преобразувател с преходи при неуспех и стандартните конструкции за HCLG претеглен преобразувател.

[4] Mihov, S., Schulz, K.U.
F-transducers for contextual text rewriting
(2021) Journal of Automata, Languages and Combinatorics, (in print)
ISSN: 1430189X; 25673785
ABSTRACT: Rule-based text rewriting is a form of language processing where a given input text/string is rewritten to a new output form using some form of knowledge base plus contextual rules. Contextual rewrite dictionaries, to be defined below, represent a general input format for such knowledge bases with contextual rules. We present an algorithm for translating a given contextual rewrite dictionary into an f-transducer. This transducer can be applied to input texts, realizing the intended form of text rewriting. F-transducers are deterministic transducers that use failure transitions in order to reduce the size. A second algorithm is given for composing f-transducers. Using composition, cascaded forms of text rewriting can be realized in a single run over the input text. Put together, a general system for rule-based text rewriting is obtained. The strength of the approach relies on four points. First, the translation algorithm is highly optimized: using the algorithm even huge knowledge bases with contextual conditions can be efficiently converted into f-transducers. Second, since f-tranducers act in a deterministic way a very efficient form of text rewriting is obtained. Third, f-transducers can also be directly applied to text streams. Fourth, composition and the general format of contextual rewrite dictionaries guarantee high flexibility as to the forms of text (stream) rewriting that can be realized.
РЕЗЮМЕ: Презаписването на текст, базирано на правила, е форма на езикова обработка, при която даден входен текст / низ се презаписва в нова изходна форма, използвайки някаква форма на база от знания плюс контекстни правила. Речниците за контекстно презаписване, които ще бъдат дефинирани по-долу, представляват общ входен формат за такива бази знания с контекстни правила. Представяме алгоритъм за преобразуване на даден контекстен речник за презаписване във f-преобразувател. Този преобразувател може да се приложи към входни текстове, реализирайки желаната форма за пренаписване на текст. F-преобразувателите са детерминирани подпоследователни преобразуватели, при които се използват преходи при неуспех, за намаляване на размера. Представен е втори алгоритъм за композиране на f-

преобразуватели. Използвайки композиция, каскада от презаписвания на текст може да бъде реализирана с едно преминаване над входния текст. Двата алгоритъма взети заедно образуват обща система за презаписване на текст базирано на правила. Преимуществото на този подход се основава на четири точки. Първо, алгоритъмът за превод е силно оптимизиран: с помощта на алгоритъма дори огромни бази от знания с контекстни условия могат да бъдат ефективно преобразувани във f-преобразуватели. Второ, тъй като f-преобразувателите действат по детерминиран начин, се получава много ефективна форма на пренаписване на текст. Трето, f-преобразувателите могат също да бъдат директно приложени към текстови потоци. Четвърто, композицията и общият формат на речниците за контекстно презаписване гарантират висока гъвкавост по отношение на формите на презаписване на текст (поток), които могат да бъдат реализирани.

[5] Mitankin, P., Mihov, S., Tinchev, T.
Large vocabulary continuous speech recognition for Bulgarian
(2009) International Conference Recent Advances in Natural Language Processing, RANLP, pp. 246-250.
ISSN: 13138502
ABSTRACT: The paper presents the results of a project completed by the authors for realizing a continuous speech recognition system for Bulgarian. The state-of-the-art speech recognition technology used in the system is discussed. Special attention is given to the problems with some specifics of the Bulgarian language namely the large vocabulary (450000 wordforms). Some implementation details of the language module are given. At the end the paper provides evaluation of the accuracy of recognition.
РЕЗЮМЕ: Докладът представя резултатите от проект, завършен от авторите за реализиране на система за разпознаване на непрекъсната реч за български език. Обсъдена е съвременната технология за разпознаване на реч, използвана в системата. Специално внимание е обърнато на проблемите с някои специфики на българския език, като например големият речник (450000 словоформи). Дадени са някои подробности за реализирането на езиковия модул. В края на статията се представят измервания на прецизността на разпознаване.

[6] Mitankin, P., Mihov, S.
A new method for real-time lattice rescoring in speech recognition
(2016) Studies in Computational Intelligence, 648, pp. 283-292.
ISSN: 1860949X
ISBN: 9783319322063
ABSTRACT: We introduce a novel efficient method, which improves the performance of speech recognition systems by providing the option to partially compile the word lattice into a deterministic finite-state automaton, making it suitable for the rescoring step in the speech recognition process. In contrast to the widely used n-best method our method permits the consideration of significantly larger number of alternatives within the same time-constraint and thus provides better recognition results. In this paper we present a description of the new method and empirical evaluation of its performance in comparison with the n-best method. The achieved WER reduction is up to 3.77% at a p-value below 3%. An important advantage of our method is its applicability for real-time applications.
РЕЗЮМЕ: Представяме нов ефективен метод, който подобрява производителността на системите за разпознаване на реч, като предоставя възможност за частично компилиране на решетката на думата в детерминиран краен автомат. Това представяне е особено подходящо за реализиран на стъпката за преоценяване на решетката в процеса на разпознаване на речта. За разлика от широко използвания метод n-best, нашият метод позволява разглеждането на значително по-голям брой алтернативи в

рамките на същото времево ограничение и по този начин осигурява по-добри резултати при разпознаването. В тази статия представяме описание на новия метод и емпирична оценка на неговото представяне в сравнение с метода n-best. Постигнатото намаление на WER е до 3,77% при p-стойност под 3%. Важно предимство на нашия метод е неговата приложимост в реално време.

[7] Hateva, N., Mitankin, P., Mihov, S.
BulPhonC: Bulgarian speech corpus for the development of ASR technology
(2016) Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016, pp. 771-774.
ISBN: 9782951740891
ABSTRACT: In this paper we introduce a Bulgarian speech database, which was created for the purpose of ASR technology development. The paper describes the design and the content of the speech database. We present also an empirical evaluation of the performance of a LVCSR system for Bulgarian trained on the BulPhonC data. The resource is available free for scientific usage.
РЕЗЮМЕ: В тази статия представяме база данни от българска реч, която е създадена с цел развитие на технологията ASR. Докладът описва дизайна и съдържанието на речевата база данни. Представяме също така емпирична оценка на презизността на система за разпознаване на непрекъсната реч с голям речник за български език, обучена с базата на BulPhonC. Ресурсът е достъпен безплатно за научна употреба.

[8] Geneva, D., Shopov, G., Mihov, S.
Building an ASR Corpus Based on Bulgarian Parliament Speeches
(2019) Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11816 LNAI, pp. 188-197.
ISSN: 03029743
ISBN: 9783030313715
ABSTRACT: This paper presents the methodology we applied for building a new corpus of Bulgarian speech suitable for training and evaluating modern speech recognition systems. The Bulgarian Parliament ASR (BG-PARLAMA) corpus is derived from the recordings of the plenary sessions of the Bulgarian Parliament. The manually transcribed texts and the audio data of the speeches are processed automatically to build an aligned and segmented corpus. NLP tools and resources for Bulgarian are utilized for the language specific tasks. The resulting corpus consists of 249 hours of speech from 572 speakers and is freely available for academic use. First experiments with an ASR system trained on the BG-PARLAMA corpus have been conducted showing word error rate of around 7% on parliament speeches from unseen speakers using time-delay deep neural network (TD-DNN) architecture. The BG-PARLAMA corpus is to our knowledge the largest speech corpus currently available for Bulgarian.
РЕЗЮМЕ: Тази статия представя методологията, която приложихме за изграждане на нов корпус от българска реч, подходящ за обучение и оценка на съвременни системи за разпознаване на реч. Корпусът на българския парламент за автоматично разпознаване на реч (BG-PARLAMA) е извлечен от записите на пленарните сесии на българския парламент. Стенограмите и аудио записите от речите се обработват автоматично, за да се изгради подравнен и сегментиран корпус. За реализирането на езиковите обработки са използвани инструментите за обработка на естествен език и ресурси за български. Полученият корпус се състои от 249 часа реч от 572 диктори и е достъпен за академична употреба. Проведени са първите експерименти с система за автоматично разпознаване на реч, обучена в корпуса на BG-PARLAMA, показваща грешка на ниво думи от около 7% в парламентарни речи от нови диктори, като се използва архитектура с дълбоки невронни мрежа със закъснение във времето (TD-

DNN). Корпусът BG-PARLAMA е, доколкото ни е известно, най-големият корпус от реч, който в момента се предлага за български език.

[9] Ringlstetter, C., Schulz, K.U., Mihov, S.
Orthographic errors in Web pages: Toward cleaner Web corpora
(2006) Computational Linguistics, 32 (3), pp. 295-340.
ISSN: 08912017
ABSTRACT: Since the Web by far represents the largest public repository of natural language texts, recent experiments, methods, and tools in the area of corpus linguistics often use the Web as a corpus. For applications where high accuracy is crucial, the problem has to be faced that a non-negligible number of orthographic and grammatical errors occur in Web documents. In this article we investigate the distribution of orthographic errors of various types in Web pages. As a by-product, methods are developed for efficiently detecting erroneous pages and for marking orthographic errors in acceptable Web documents, reducing thus the number of errors in corpora and linguistic knowledge bases automatically retrieved from the Web.
РЕЗЮМЕ: Интернет представлява най-голямата публична база от текстове на естествен език, поради което скорошните експерименти, методи и инструменти в областта на корпусната лингвистика често използват Интернет мрежата като корпус. За приложения, при които висока правописна коректност е от решаващо значение, възниква проблема, че в Интернет документите се появяват значителен брой правописни и граматични грешки. В тази статия ние изследваме разпространението на правописни грешки от различен тип в Интернет страниците. Като страничен продукт разработваме методи за ефективно откриване на грешки в страниците и за маркиране на правописни грешки в Интернет документи, като по този начин се намалява броят на грешките в корпусите и базите от езикови знания, автоматично извлечени от мрежата.

[10] Mihov, S., Mitankin, P., Schulz, K.U.
Fast selection of small and precise candidate sets from dictionaries for text correction tasks
(2007) Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, 1, art. no. 4378754, pp. 471-475.
ISSN: 15205363
ISBN: 0769528228; 9780769528229
ABSTRACT: Lexical text correction relies on a central step where approximate search in a dictionary is used to select the best correction suggestions for an ill-formed input token. In previous work we introduced the concept of a universal Levenshtein automaton and showed how to use these automata for efficiently selecting from a dictionary all entries within a fixed Levenshtein distance to the garbled input word. In this paper we look at refinements of the basic Levenshtein distance that yield more sensible notions of similarity in distinct text correction applications, e.g. OCR. We show that the concept of a universal Levenshtein automaton can be adapted to these refinements. In this way we obtain a method for selecting correction candidates which is very efficient, at the same time selecting small candidate sets with high recall.
РЕЗЮМЕ: Лексикалната корекция на текст разчита на централна стъпка, която използва приближено търсене в речник, за да се изберат най-добрите предложения за корекция за грешните думи. В предишна работа представихме концепцията за универсален Левенщайн автомат и показахме как да използваме тези автомати за ефективно намиране от речник на всички думи на фиксирано разстояние по Левенщайн до грешната входна дума. В тази статия разглеждаме усъвършенстванията на основното разстояние на Левенщайн, които дават по-релевантни разстояния за

близост в различни приложения за корекция на текст, например след OCR. Ние показваме, че концепцията за универсален автомат на Левенщайн може да бъде адаптирана към тези усъвършенствания. По този начин получаваме метод за подбор на кандидати за корекция, който е много ефективен, като в същото време намираме малки набори от кандидати с високо покритие.

[11] Mihov, S., Mitankin, P., Gotscharek, A., Reffle, U., Schulz, K.U., Ringlstetter, C.
Using automated error profiling of texts for improved sgelection of correction candidates for garbled tokens
(2007) Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 4830 LNAI, pp. 456-465.
ISSN: 03029743
ISBN: 9783540769262
ABSTRACT: Lexical text correction systems are typically based on a central step: when finding a malformed token in the input text, a set of correction candidates for the token is retrieved from the given background dictionary. In previous work we introduced a method for the selection of correction candidates which is fast and leads to small candidate sets with high recall. As a prerequisite, ground truth data were used to find a set of important substitutions, merges and splits that represent characteristic errors found in the text. This prior knowledge was then used to fine-tune the meaningful selection of correction candidates. Here we show that an appropriate set of possible substitutions, merges and splits for the input text can be retrieved without any ground truth data. In the new approach, we compute an error profile of the erroneous input text in a fully automated way, using socalled error dictionaries. From this profile, suitable sets of substitutions, merges and splits are derived. Error profiling with error dictionaries is simple and very fast. As an overall result we obtain an adaptive form of candidate selection which is very efficient, does not need ground truth data and leads to small candidate sets with high recall.
РЕЗЮМЕ: Системите за лексикална корекция на текст обикновено се основават на централна стъпка: при намиране на сгрешен низ във входния текст, набор от кандидати за корекция на низа се извлича от даден базов речник. В наша предишна работа представихме метод за избор на кандидати за корекция, който е бърз и води до малък набор от кандидати с високо покритие. Като предпоставка беше използвана база от ръчно поправени текстове, за да се намери набор от заместванията, сливания и разделяния на символи, които представляват характерни грешки, срещани в текстовете. Тези предварителни знания бяха използвани за по-точен подбор на релевантните кандидати за корекция. Тук показваме, че набор от заместванията, сливания и разделяния на символи, които представляват характерни грешки могат да бъдат извлечени без никакви предварително поправени текстове. В новия подход ние изчисляваме профила на грешките на грешния въведен текст по напълно автоматизиран начин, като използваме така наречените речници за грешки. От този профил се извличат подходящи набори от заместванията, обединения и разделяния. Профилирането на грешки с речници за грешки е лесно и много бързо. Като общ резултат получаваме адаптивна форма на подбор на кандидати, която е много ефективна, не се нуждае от предварително поправени данни и води до малки набори от кандидати с високо покритие.

[12] Ringlstetter, C., Hadersbeck, M., Schulz, K.U., Mihov, S.
Text Correction Using Domain Dependent Bigram Models fromWeb Crawls
(2007) Proceedings of IJCAI 2007 Workshop on Analytics for Noisy Unstructured Text Data, AND 2007, pp. 47-54.
ISSN: 15504875

ABSTRACT: The quality of text correction systems can be improved when using complex language models and by taking peculiarities of the garbled input text into account. We report on a series of experiments where we crawl domain dependent web corpora for a given garbled input text. From crawled corpora we derive dictionaries and language models, which are used to correct the input text. We show that correction accuracy is improved when integrating word bigram frequency values from the crawls as a new score into a baseline correction strategy based on word similarity and word (unigram) frequencies. In a second series of experiments we compare the quality of distinct language models, measuring how closely these models reflect the frequencies observed in a given input text. It is shown that crawled language models are superior to language models obtained from standard corpora.
РЕЗЮМЕ: Качеството на системите за корекция на текст може да се подобри, като се използват сложни езикови модели и като се вземат предвид особеностите на входния текст. Ние представяме поредица от експерименти, при които извличаме автоматично корпуси от интернет в зависимост от областта на дадения входен текст. От интернет корпусите извличаме речници и езикови модели, които се използват за коригиране на входния текст. Ние показваме, че точността на корекцията се подобрява при добавяне на честотата на биграмите на думите от интернет корпусите като нов фактор в базовата стратегия за корекция основаваща се на близостта и униграмната честота на думите. Във втора поредица от експерименти сравняваме качеството на различните езикови модели, измервайки колко точно тези модели отразяват честотите, наблюдавани в даден входен текст. Показваме, че езиковите модели от интернет корпуси превъзхождат езиковите модели, получени от стандартните корпуси.

[13] Mihov, S., Schulz, K.U.
Computation of similarity-similarity search as computation
(2011) Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 6735 LNCS, pp. 201-210.
ISSN: 03029743
ISBN: 9783642218743
ABSTRACT: We present a number of applications in Natural Language Processing where the main computation consists of a similarity search for an input pattern in a large database. Afterwards we describe some efficient methods and algorithms for solving this computational challenge. We discuss the view of the similarity search as a special kind of computation, which is remarkably common in applications of Computational Linguistics.
РЕЗЮМЕ: Представяме редица приложения в обработката на естествен език, където основното изчисление се състои в търсене на близост на входен шаблон в голяма база данни. След това описваме някои ефективни методи и алгоритми за решаване на този изчислителен проблем. Ние разглеждаме търсенето на близост като специален вид изчисление, който е изключително разпространен в приложенията на компютърната лингвистика.

[14] Gerdjikov, S., Mihov, S., Mitankin, P., Schulz, K.U.
Good parts first - a new algorithm for approximate search in lexica and string databases
(2013) arXiv:1301.0722
(2015) arXiv:1301.0722v2
ABSTRACT: We present a new efficient method for approximate search in electronic lexica. Given an input string (the pattern) and a similarity threshold, the algorithm retrieves all entries of the lexicon that are sufficiently similar to the pattern. Search is organized in subsearches that always start with an exact partial match where a substring of the input pattern is aligned with a substring of a lexicon word. Afterwards this partial match is extended stepwise to larger substrings. For aligning further parts of the pattern with corresponding parts of lexicon entries, more errors are tolerated at each subsequent step. For

supporting this alignment order, which may start at any part of the pattern, the lexicon is represented as a structure that enables immediate access to any substring of a lexicon word and permits the extension of such substrings in both directions. Experimental evaluations of the approximate search procedure are given that show significant efficiency improvements compared to existing techniques. Since the technique can be used for large error bounds it offers interesting possibilities for approximate search in special collections of "long" strings, such as phrases, sentences, or book titles.
РЕЗЮМЕ: Представяме нов ефективен метод за приближено търсене в електронен речник. При даден входен низ (шаблон) и праг за близост, алгоритъмът извлича всички записи в речника, които са достатъчно близки с шаблона. Търсенето е организирано в подтърсения, които винаги започват с точно частично съвпадение, където подниз от входния шаблон е подравнен с подниз от речникова дума. След това това частичното съвпадение се разширява постепенно до по-големи поднизове. За подравняване на по-нататъшни части на шаблона със съответните части на речниковите записи се допускат повече грешки на всяка следваща стъпка. За поддържане на този ред на подравняване, който може да започне от която и да е част на шаблона, речникът е представен като структура, която позволява директен достъп до всеки подниз на речникова дума и позволява разширяването на такива поднизове в двете посоки. Дадени са експериментални оценки на процедурата за приближено търсене, които показват значително подобрение на ефективността в сравнение със съществуващите техники. Тъй като техниката може да се използва за големи прагове за близост, тя предлага интересни възможности за приближено търсене в специални колекции от "дълги" низове, като фрази, изречения или заглавия на книги.

[15] Gerdjikov, S., Mihov, S., Mitankin, P., Schulz, K.U.
WallBreaker - Overcoming the wall effect in similarity search
(2013) ACM International Conference Proceeding Series, pp. 366-369.
ISBN: 9781450315999
ABSTRACT: In this paper we present the WallBreaker system for similarity search as used in the String Similarity Search/Join Competition, 2013, organized by the Humboldt University of Berlin. We consider the problem of how to efficiently find for a given string P (pattern) all words W in a lexicon such that the distance between P and W does not exceed a given bound b. Classical solutions to this problem try to align P with suitable lexicon words in a strict left-to-right manner, starting at the left border of the pattern. During the search, only prefixes of lexicon words are visited where the distance to a prefix P' of the pattern does not exceed the given bound b. The main problem with this solution is the so-called "wall effect": if we tolerate b errors and start searching in the lexicon from left to right, then in the first b steps we have to consider all prefixes of lexicon words. Eventually, only a tiny fraction of these prefixes will lead to a useful lexicon word, which means that our exhaustive initial search represents a waste of time. To avoid the "wall effect", in WallBreaker we have implemented our new method. To sketch it let us assume that the pattern can be aligned with a lexicon word with not more than b errors. Clearly, if we divide the pattern into b+1 pieces, then at least one piece will exactly match the corresponding substring of a lexicon word in the answer set. In our approach we first find the lexicon substrings that exactly match such a given piece of the pattern. Afterwards we continue by extending this alignment, step-wise attaching new pieces on the left or right side. For the alignment of new pieces, more errors are tolerated at each step, which guarantees that eventually b errors can occur. Since at later steps the set of interesting substrings to be extended is already small the wall effect is avoided, it does not hurt that we need to tolerate more errors. For this kind of search strategy, a new representation of the lexicon is needed where we can start traversal at any point of a word. In our new approach, the lexicon is represented as symmetric compact directed acyclic word graph (SCDAWG). This index

structure can be seen as a part of a longer development of related index structures. Our implementation executes the search queries in parallel. It is realized in ANSI C, compiled with GCC and does not use any additional libraries beside LIBC and POSIX threads. In average it performs a similarity search of a 100 character pattern with up to 16 errors in a lexicon with 750 000 entries in about 0.088 ms.
РЕЗЮМЕ: В тази статия представяме системата WallBreaker за търсене на близост, използвана в състезанието за търсене на близост на низове, 2013, организирано от университета Хумболт в Берлин. Ние разглеждаме задачата за това как да намерим ефективно за даден низ P (шаблон) всички думи W в речник, така че разстоянието между P и W да не надвишава дадена граница b. Класическите решения на този проблем се опитват да подравнят P с подходящи речникови думи стриктно отляво надясно, започвайки от ляво на шаблона. По време на търсенето се посещават само префикси от речникови думи, при които разстоянието до префикс P' на шаблона не надвишава дадената граница b. Основният проблем с това решение е така нареченият „ефект на стената“: ако толерираме b грешки и започнем да търсим в речника отляво надясно, тогава в първите b стъпки трябва да разгледаме всички префикси на речникови думи. В крайна сметка само малка част от тези префикси ще доведат до полезна речникова дума, което означава, че нашето изчерпателно първоначално търсене представлява загуба на време. За да избегнем "ефекта на стената", в WallBreaker внедрихме нашия нов метод. За да го скицираме, нека приемем, че шаблонът може да бъде подравнен с речникова дума с не повече от b грешки. Ясно е, че ако разделим шаблона на b+1 парчета, тогава поне едно парче ще съвпада със съответния подниз на речниковата дума. В нашия подход първо намираме речниковите поднизове, които съвпадат с дадено парче от шаблона. След това продължаваме, като разширяваме това подравняване, като на всяка стъпка прикрепяме нови парчета от лявата или дясната страна. За подравняването на нови парчета се допускат повече грешки на всяка стъпка, което гарантира, че в крайна сметка може да възникнат b грешки. Тъй като в по-късните стъпки наборът от интересни поднизове, които трябва да се разшири, вече е малък, ефектът на стената се избягва и не е проблем, че трябва да толерираме повече грешки. За този вид стратегия за търсене е необходимо ново представяне на речника, което позволява да започнем обхождане във всяка точка на думата. В нашия нов подход речника е представен като симетричен компактен насочен ацикличен граф от думи (SCDAWG). Тази индексна структура може да се разглежда като резултат от развитието на индексни структури в тази област. Нашата програма изпълнява паралелно заявките за търсене. Тя е реализиран на ANSI C, компилирана с GCC и не използва никакви допълнителни библиотеки освен LIBC и POSIX нишки. Програмата средно извършва търсене на близост на шаблон от 100 знака с до 16 грешки в речник със 750 000 записа за около 0,088 ms.

[16] Gerdjikov, S., Mihov, S., Nenchev, V.
Extraction of spelling variations from language structure for noisy text correction
(2013) Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, art. no. 6628637, pp. 324-328.
ISSN: 15205363
ABSTRACT: We describe a novel approach for the extraction of spelling variations from a list of instances. It relates distinctive infixes to distinctive infixes of referenced words. The distinctive infixes are extracted automatically from a (multi)set of instances and a referenced dictionary without any additional expert knowledge. Based on the spelling variations retrieved during a learning(training) phase we develop a correction algorithm which suggests and ranks candidates for a particular noisy word. The main advantage of our approach is that it provides good corrections for the unobserved noisy words while it is almost perfect on words observed during the learning. Our experimental results of the

normalisation of a typical reference corpus of Early Modern English letters, significantly improve over previous results of VARD2. We also achieve better results than those reported on the OCR-correction of the TREC-5 Confusion Track corpus.
РЕЗЮМЕ: Описваме нов подход за извличане на вариации на правописа от списък с примери. Той свързва отличителни инфикси с отличителни инфикси на референтни думи. Отличителните инфикси се извличат автоматично от (мулти) множество от примери и референтен речник без никакви допълнителни експертни знания. Въз основа на вариациите на правописа, получени по време на фаза на обучение, ние разработихме алгоритъм за корекция, който предлага и ранкира кандидатите за определена грешна дума. Основното предимство на нашия подход е, че той осигурява добри корекции за нови грешни думи, като е почти перфектен за думите, наблюдавани по време на обучението. Нашите експериментални резултати от нормализирането на типичен референтен корпус от писма на ранен модерен английски значително подобряват резултатите постигнати от ситемата VARD2. Също така постигаме по-добри резултати от тези, докладвани за OCR-корекция на корпуса от TREC-5 Confusion Track.

[17] Sariev, A., Nenchev, V., Gerdjikov, S., Mitankin, P., Ganchev, H., Mihov, S., Tinchev, T.
Flexible noisy text correction
(2014) Proceedings - 11th IAPR International Workshop on Document Analysis Systems, DAS 2014, art. no. 6830964, pp. 31-35.
ISBN: 9781479932436
ABSTRACT: We present a new general and language independent approach to the noisy text correction problem developed and implemented in the framework of the CULTURA project. We briefly describe the core candidate generator, REBELS, the complete system concept, its efficient implementation based on functional automata and its immediate applications. The quality of the whole system is empirically established in different experimental settings where language and noise sources are varied.
РЕЗЮМЕ: Представяме нов общ и езиковонезависим подход за корекция на текстове, разработен и имплементиран в рамките на проекта CULTURA. Накратко описваме основния генератор на кандидати на думи за корекция, REBELS, цялостната идея на системата, нейната ефективна имплементация, основаваща се на функционални автомати и нейното непосредствено приложение. Качеството на цялата система е оценено в различни тестови установки, при които се мени езикът и произходът на грешки/шум.

[18] von Groll; G, Mihov; S, Solheim; C, Dyrdal; D,
Media-based computational influencer network analysis,
(2011) United States Patent 7,933,843
ABSTRACT: The methodology draws from three disciplines, namely public relations, social network analysis and computer-based information extraction. The analysis permits the visualization of how various people, organizations, products, subjects, key messages etc. are linked/form a network dynamic in media coverage. This type of analysis can assist corporations and other organizations to understand, plan and measure the effectiveness of communication.
РЕЗЮМЕ: Методологията комбинира резултати от три области, а именно връзки с обществеността, анализ на социални мрежи и компютърно извличане на информация. Анализът позволява визуализиране на това как различни хора, организации, продукти, теми, ключови съобщения и т.н. са свързани / формират мрежова динамика в медийното отразяване. Този тип анализ може да помогне на корпорациите и други организации да разберат, планират и измерват ефективността на комуникацията.