и пропускать обрабатываемый текст последовательно через несколько грамматик и словарей в заданном порядке, таким образом добиваясь на каждой итерации своих результатов. Результатом работы парсера являются выделенные с помощью грамматик и словарей слова и словоцепочки, а также граф предложений, который впоследствии можно преобразовать в единый текстовый граф слов, и с помощью этого графа можно будет выделить знания и данные, которые относятся к ключевому или выделяемому слову, соотнести факты относящиеся к одному понятию и произвести сооружение дополнительных связей на основе этого графа.

На основе вышеизложенного можно сделать вывод, что Томита-парсер, с такими произведенными улучшениями, как агрегатор графов предложений в общий граф слов и грамматики, описывающие наиболее характерные для данного текста цепочки является мощным инструментом в обработке текстовой информации.

Список использованных источников:

1. [Электронный источник] Технологии Яндекса — Томита-парсер. https://tech.yandex.ru/tomita/ Дата доступа 10.03.2015 г.
2. [Электронный источник] GitHub. GLR-парсер. https://github.com/vas3k/python-glr-parser Дата доступа 10.03.2015 г.

# АЛГАРЫТМ АЎТАМАТЫЧНАГА ВЫЗНАЧЭННЯ МЕСЦА НАЦІСКУ Ў НЕВЯДОМЫХ СЛОВАХ У ЛІНГВІСТЫЧНАЙ ІНФАРМАЦЫЙНА-ПОШУКАВАЙ СІСТЭМЕ

*Беларускі дзяржаўны ўніверсітэт інфарматыкі і радыёэлектронікі*
*г. Мінск, Рэспубліка Беларусь*

*Філіпчык А. В.*

*Сярэбраная Л. В. — к. т. н., дацэнт*

Разгледжаныя асаблівасці вызначэння націску. Прапанаваны алгарытм, які падыходзіць для рашэння пастаўленай задачы вызначэння націскнога складу ў невядомых словах.

Для многіх сістэм, якія працуюць з тэкставымі дадзенымі, неабходна ведаць не толькі правільнае напісанне слова, але яшчэ і правільнае яго вымаўленне, у прыватнасці, трэба ведаць месца націску. Напрыклад, такія дадзеныя вельмі важныя для сістэм сінтэзу маўлення, альбо для камерцыйных прадуктаў тыпу Soundex, што прымяняецца авіякампаніямі для захоўвання імён і прозвішчаў пасажыраў у фанетычнай форме для прадухілення канфліктных сітуацый з няправільным напісаннем ідэнтыфікацыйных дадзеных у розных мовах і алфавітах. Звычайна такія сістэмы аперыруюць слоўнікамі, у якіх захоўваецца ўся неабходная інфармацыя пра кожную лексічную адзінку, але натуральныя мовы імкліва развіваюцца, і ні адзін слоўнік не можа ўмясціць абсалютна ўсе славаформы, якія могуць сустрэцца сістэме. Для вызначэння націску ў невядомых словах неабходны адмысловыя алгарытмы, гэтаму і прысвечана дадзеная праца.

На ўваход распрацаванага алгарытма могуць паступаць як асобныя словы, так і цэлыя тэксты.

Алгарытм вызначэння націску можна падзяліць на наступныя этапы:

1. Марфалагічны аналіз і пошук слова ў базе дадзеных.
2. Вызначэнне стандартных прэфіксаў.
3. Вызначэнне стандартных суфіксаў і канчаткаў.
4. Прадказанне націску на аснове статыстыкі.

На першым этапе алгарытм праводзіць марфалагічны аналіз кожнай лексічнай адзінкі для выяўлення выпадкаў амаграфіі. Відавочна, што зварот да слоўніка не заўсёды дазваляе адназначна вызначыць прыналежнасць славаформы да той ці іншай лексемы. Амаграфы – гэта славаформы з аднолькавым напісаннем, якія, тым не менш, належаць да розных лексем і могуць адрознівацца націскам. Метад кантэкстнага аналіза ўлічвае славаформы ў левым і правым кантэксце і падлічвае імавернасць з'яўлення ў дадзеным кантэксце той ці іншай граматычнай формы. Пасля марфалагічнага аналізу алгарытм шукае слова ў базе славаформ па вызначаных характарыстыках. База славаформ пабудавана на аснове слоўнікаў беларускай мовы, узятых з адкрытых крыніц (усяго больш за 500.000 уваходжанняў). Аднак у гэтыя слоўнікі ўключаныя далёка не ўсе існыя словы: так, у іх адсутнічаюць шматлікія імёны, назвы, рэгіяналізмы, аўтарскія словы, неалагізмы. Між тым, алгарытм павінны з высокай імавернасцю правільна вызначаць націск ва ўсіх словах.

Другі этап - вызначэнне стандартных прэфіксаў. У аснове метада вызначэння націска - формула $p = (n+1)/2$, дзе n – колькасць складоў у слове. Гэты алгарытм дае добрыя вынікі ў кароткіх словах, але ў доўгіх словах, асабліва складаных, дае даволі вялікую колькасць памылак. Для паляпшэння вынікаў выкарыстоўваецца механізм вылучэння стандартных прэфіксаў. У спіс прэфіксаў уваходзяць як прыстаўкі, так і першыя часткі складаных словаў (такія як пяцьсот-, трактара-, стара-, электра- і г.д). Слова дзеліцца на дзве часткі: прэфікс і аснову, якая зноў шукаецца ў слоўніку. Некаторыя прэфіксы маюць уласны націск,

але большая частка яго не мае. У любым выпадку, вызначэнне стандартнага прэфікса спрашчае задачу алгарытма.

Трэці этап – вызначэнне стандартных суфіксаў і канчаткаў. Для гэтага аналізуецца канцавы сегмент слова. Праводзіцца пошук суфіксаў, на якія звычайна падае націск (-ятк-, -еньк-), а таксама ненаціскных суфіксаў (напрыклад, паслянаціскны -нік). Акрамя таго, разглядаюцца другія часткі складаных словаў (-здольны) і стандартных частак спецыфічных прозвішчаў (прозвішчы на -швілі і -адзэ).

Апошні этап – прадказанне націску на аснове статыстыкі. У аснове механізма – гіпотэза аб тым, што ў большасці выпадкаў можна вызначыць месца націску ў слове па яго канцавой частцы. Алгарытм працуе з загадзя пабудаваным слоўнікам на аснове слоў з вядомым націскам з базы дадзеных. Гэты слоўнік утрымлівае канцавыя часткі слоў і “правілы”, якія вызначаюць націск для ўваходнага слова на аснове яго марфалагічнай прыналежнасці і колькасці складоў. Эксперыментальным шляхам было вызначана, што даўжыня канцавой часткі павінна быць не менш за 4 сімвалы, а працэнт слоў з найбольш імаверным месцам націску – не менш за 80. Так, напрыклад, для 4-складовых словаў з канцавой часткай –лава у 93% выпадкаў націск ставіцца на перадапошні склад.

Атрыманы алгарытм, які дазваляе з высокай імавернасцю вызначаць націскны склад у незнаёмых словах, быў распрацаваны ў межах стварэння лінгвістычнай інфармацыйна-пошукавай сістэмы для аўтаматызацыі падбору рыфм і напісання вершаваных радкоў. Таксама дадзены алгарытм можа быць выкарыстаны ў сістэмах аўтаматычнага сінтэзу маўлення, аналізатарах галасавых каманд і іншых сферах.

Спіс выкарыстаных крыніц:

1. Кипяткова, И.С. Разработка и оценивание модуля транскрибирования для распознавания и синтеза русской речи / И.С. Кипяткова, А.А. Карпов. - Научно-теоретический журнал "Искусственный интеллект" No.3'2009.
2. Кондратов, А. Математика и поэзия. /А. Кондратов - М.: Знание, 1962. - 42 с.
3. Слоўнік.орг [Электронны рэсурс]. – 2015 – Рэжым доступу: http://www.slounik.org / Дата доступу : 22.02.2015

# ПРОГРАММНОЕ СРЕДСТВО УПРАВЛЕНИЯ РАСПРЕДЕЛЕНИЕМ ДАННЫХ МЕЖДУ ОБЛАЧНЫМИ ХРАНИЛИЩАМИ

*Белорусский государственный университет информатики и радиоэлектроники*
*г. Минск, Республика Беларусь*

*Гавриленко Д.В.*

*Куликов С. С. – к.т.н., доцент*

Современный этап развития общества подразумевает эффективное управление данными, и в данном контексте особую важность приобретают такие системы, в которых данные хранятся на многочисленных распределённых в сети серверах – так называемых облачных хранилищах данных.

В противовес модели хранения данных на собственных (выделенных) серверах, приобретаемых или арендуемых специально для нужд функционирования системы, внутренняя инфраструктура облачного хранилища может быть неизвестна потребителю услуг: данные хранятся и обрабатываются в так называемом «облаке», которое представляет собой с точки зрения клиента один большой виртуальный сервер [1].

Облачные шлюзы – технология, которая может быть использована для более удобного представления облака клиенту и более гибкой настройки многопользовательских высоконагруженных систем. К примеру, с помощью соответствующего программного обеспечения хранилище в облаке может быть представлено для клиента как локальный диск на компьютере. Таким образом, работа с данными в облаке для клиента становится абсолютно прозрачной. И при наличии хорошей, быстрой связи с облаком клиент может даже не замечать, что работает не с локальными данными у себя на компьютере, а с данными, хранящимися, возможно, за много сотен километров от него.

Одним из важных вопросов является безопасность передаваемых и хранимых данных, поэтому важно, чтобы облачный сервис использовал надёжные протоколы шифрования, что на сегодняшний день в полной мере поддерживается большинством стандартных решений в области облачного хранения данных [2].

Другим важным показателем качества облачного хранилища является надёжность хранения и доступность данных в облаке: здесь стоит отметить, что основную проблему для отечественных потребителей услуг облачного хранения данных представляют некачественные и медленные каналы передачи данных, т.к. сами по себе облачные хранилища обеспечивают в среднем более высокую надёжность хранения данных, чем локальные решения, что достигается за счёт многократного нагруженного резервирования, распределения данных между несколькими узлами хранения и иных технологических решений [3].

Кроме того, не стоит полностью доверять хранение важных документов какому-либо одному сервису, так как известны случаи удаления файлов пользователей по инициативе владельцев сервисов, например, если появились основания полагать, что файл нарушает авторские права. Все это наводит на мысли некоторого объединения нескольких облачных сервисов с использованием технологий, похожих на