Міністэрства адукацыі Рэспублікі Беларусь
Установа адукацыі
Беларускі дзяржаўны ўніверсітэт
Інфарматыкі і радыёэлектронікі

УДК 004.657

Філіпчык
Андрэй Васільевіч

Мадэлі, метады, праграмны сродак аналізу і пабудовы лінгвістычнай інфармацыйна-пошукавай сістэмы

АЎТАРЭФЕРАТ
на саісканне акадэмічнай ступені
магістра тэхнічных навук

па спецыяльнасці 1-40 80 05 – Матэматычнае і праграмнае забеспячэнне вылічальных машын, комплексаў і камп'ютарных сетак

Навуковы кіраўнік
Сярэбраная Л.В.
к.т.н., дацэнт

Мінск 2015

КАРОТКІЯ ЎВОДЗІНЫ

У апошнія дзесяцігоддзі стала відавочна, што галоўная каштоўнасць у сучасным свеце – гэта інфармацыя. Таму надзвычай важнымі сталі сістэмы, якія дазваляюць гэтую інфармацыю знайсці. Інфармацыйна-пошукавыя сістэмы (ІПС) з'явіліся як вынік развіцця аўтаматызацыі захоўвання і пошуку інфармацыі, задач, якія раней маглі выконвацца толькі чалавекам.

На сёння ІПС знайшлі шырокае прымяненне практычна ва ўсіх галінах, дзе наогул магчымая аўтаматызацыя: у прамысловасці, банкаўскай справе, гандлі, лінгвістыцы. Аднак, не гледзячы на разнастайнасць сфер прымяненне, у аснову ўсіх ІПС закладзеныя адныя і тыя ж базавыя прынцыпы. Пераўтварэнне інфармацыі ў ІПС зводзіцца не толькі да маніпулявання дадзенымі, але і да апрацоўкі ведаў, выражаных у тэрмінах канкрэтнай прадметнай галіны.

Шматлікія аўтаматызаваныя сістэмы сутыкаюцца з неабходнасцю аналізу тэкставых дадзеных на натуральных мовах. Першая, самая нізкая прыступка пры гэтым аналізе – гэта фанетыка (аналіз гукаў). Потым ідзе марфалогія, якая вывучае структуру слова. Яшчэ вышэй стаіць сінтаксіс, які займаецца сувязямі паміж словамі. Для вырашэння праблем, якія паўстаюць пры аналізе такіх дадзеных, існуюць лінгвістычныя ІПС.

Лінгвістычная ІПС, якая разглядаецца ў дысертацыйнай рабоце, канцэнтруе сваю ўвагу на падборы рыфмы. З яе дапамогай можна падбіраць рыфмы да зададзеных словаў ці цэлых паэтычных радкоў. Акрамя гэтага, дадзеная ІПС мае магчымасць папаўняць свой слоўнік новымі словамі, праводзіць іх марфалагічны аналіз, разлічвае націск.

Лінгвістычныя ІПС могуць прымяняцца не толькі ў вершаскладанні, але і ў электронных перакладчыках, сістэмах сінтэзу маўлення і аналізу галасавых камандаў і гэтак далей. З часам колькасць і разнастайнасць сістэм, якія размаўляюць з карыстальнікамі жывой, натуральнай мовай, толькі расце, а значыць, павялічваецца і патрэба ў пабудове эфектыўных лінгвістычных ІПС.

АГУЛЬНАЯ ХАРАКТАРЫСТЫКА РАБОТЫ

Мэта і задачы даследвання

Мэтай дысертацыйнай работы з'яўляецца пабудова лінгвістычнай інфармацыйна-пошукавай сістэмы. Для гэтага неабходна вырашыць наступныя *задачы*:

1. Даследаваць прынцыпы пабудовы лінгвістычных інфармацыйна-пошукавых сістэм.
2. Даследаваць існуючыя мадэлі і метады аналізу і пабудовы лінгвістычных інфармацыйна-пошукавых сістэм, прааналізаваць іх перавагі і недахопы.
3. Рэалізаваць інфармацыйную мадэль пошуку дадзеных, якая будзе эфектыўна шукаць рыфму да зададзенага слова.
4. Распрацаваць эфектыўны метад стэмінгу словаў для выкарыстання ў лінгвістычнай ІПС.
5. Распрацаваць метад вызначэння націску ў словах.
6. Распрацаваць сістэму транскрыбавання словаў для максімальна дакладнай перадачы вуснага маўлення на пісьме.
7. Распрацаваць сістэму ранжыравання і класіфікацыі вынікаў пошуку.
8. Рэалізаваць праграмны сродак, які дазволіць прымяніць распрацаваныя мадэлі і метады на практыцы.

Аб'ектам даследвання з'яўляецца лінгвістычная інфармацыйна-пошукавая сістэма.

Прадметам даследвання з'яўляюцца мадэлі і метады аналіза і пабудовы лінгвістычных інфармацыйна-пошукавых сістэм.

Асноўнай *гіпотэзай*, пакладзенай у аснову дысертацыйнай работы, з'яўляецца магчымасць распрацаваць метады і алгарытмы эфектыўнай аўтаматызацыі пошуку рыфмы. Для гэтага выкарыстоўваюцца нармалізацыя словаў, транскрыпцыя і пошук націскнога складу.

Сувязь работы з прыярытэтнымі накірункамі навуковых даследванняў і запытамі рэальнага сектара эканомікі

Работа выконвалася ў адпаведнасці з навукова-тэхнічным заданнем і планам работ кафедры «Праграмнае забеспячэнне інфармацыйных тэхналогій» па тэме «Распрацаваць мадэлі, метад, алгарытмы для ацэнкі параметраў, падвышэння надзейнасці і якасці функцыянавання апаратна-праграмных сродкаў сістэм і сетак складанай канфігурацыі і ўкараніць у сучасныя навучальныя комплексы» (ГБ № 11-2004, № ГР 20111065, навуковы кіраўнік НДР – В. В. Бахцізін).

Асабісты ўнёсак саіскальніка

Вынікі, прыведзеныя ў дысертацыі, атрыманыя саіскальнікам асабіста. Унёсак навуковага кіраўніка Л.В. Сярэбранай, заключаецца ў фармуліроўцы мэты і задач даследвання.

Апрабацыя вынікаў дысертацыі

Асноўныя палажэнні дысертацыйнай работы дакладаліся і абмяркоўваліся на Рэспубліканскай навуковай канферэнцыі студэнтаў і аспірантаў «Актуальныя пытанні навукі і тэхнікі» (Гомель, Беларусь, 2015); 51-ай навуковай канферэнцыі аспірантаў, магістрантаў і студэнтаў "Камп'ютарныя сістэмы і сеткі" (БДУІР, Мінск, Беларусь, 2015).

Апублікаванасць вынікаў дысертацыі

Па тэме дысертацыі апублікавана 2 друкаваныя работы, з іх 2 работы ў зборніках прац і матэрыялаў канферэнцый.

Структура і аб'ём дысертацыі

Дысертацыя складаецца з уводзінаў, чатырох раздзелаў, заключэння, спіса выкарыстаных крыніц, спіса публікацый аўтара. У першым раздзеле дысертацыйнай работы зроблены агляд прадметнай галіны. Разгледжаная класіфікацыя, гісторыя стварэння і асноўныя прынцыпы інфармацыйна-пошукавых сістэм увогуле і лінгвістычных інфармацыйна-пошукавых сістэм у прыватнасці. Другі раздзел прысвечаны аналізу існуючых мадэляў і метадаў пабудовы лінгвістычных інфармацыйна-пошукавых сістэм: разгледжаныя іерархічная, сеткавая і рэляцыйная мадэлі захоўвання інфармацыі, а таксама булева, вектарная і імавернасная мадэлі пошуку дадзеных; разгледжаныя метады стэмінгу (нармалізацыі) словаў і вызначэння месца націску. Сфармуляваныя мэты і задачы. Трэці раздзел прысвечаны пабудове лінгвістычнай інфармацыйна-пошукавай сістэмы і ўтрымлівае даследване прынцыпаў пабудовы, пытанняў эфектыўнага захоўвання вялікіх аб'ёмаў інфармацыі, хуткага пошуку, мадыфікацыі і дадання новай інфармацыі, стварэння зручнага карыстальніцкага інтэрфейсу. У чацвёртым раздзеле апісваецца прымяненне пры стварэнні праграмнага сродку мадэляў і метадаў, асаблівасці іх працы ў дадзеным праграмным сродку. Раздзел утрымлівае мадэлі, спецыфікацыі патрабаванняў і праектаванне праграмнага сродку.

Агульны аб'ём работы складае 71 старонку, з якіх асноўнага тэксту – 58 старонак, 20 малюнкаў на 10 старонках, і спіс выкарыстаных крыніц з 30 найменняў на 3 старонках.

АСНОЎНЫ ЗМЕСТ

Ва **ўводзінах** вызначаная галіна і ўказаныя асноўныя напрамкі даследвання, паказаная актуальнасць тэмы дысертацыйнай работы, дадзеная

кароткая характарыстыка даследуемых пытанняў, абазначаная практычная каштоўнасць работы.

У **першым раздзеле** праведзены агляд прадметнай галіны. Разгледжаная гісторыя стварэння інфармацыйна-пошукавых сістэм ад 70-ых гадоў 19 стагоддзя да сучаснасці. Прыведзеная класіфікацыя і асноўныя прынцыпы інфармацыйна-пошукавых сістэм. Праведзены аналіз існуючых лінвістычных інфармацыйна-пошукавых сістэм. Сфармуляваныя мэта і задачы дысертацыйнай работы.

Сучасныя лінгвістычныя інфармацыйна-пошукавыя сістэмы характарызуюцца тым, што цесна сутыкаюцца з жывой чалавечай мовай, якая не такая стандартызаваная і схематычная як машынныя каманды і нежывыя дадзеныя. У задачы лінгвістычных ІПС можа ўваходзіць аналіз фанэтыкі (гукаў), марфалогіі (разбор і класіфікацыя словаў), сінтаксісу (аналіз сказаў), семантыкі (вывучэнне сэнсу дадзеных). Акрамя аналізу тэкстаў лігвістычныя ІПС могуць прымяняцца пры перакладзе, у сістэмах аўтаматызацыі стасункаў з карыстальнікамі, пры трансляцыі галасавых дадзеных у тэкставыя і наадварот.

Пашырэнне функцыянальных магчымасцяў ЛІПС дасягаецца праз пашырэнне і паляпшэнне базавых функцый, але таксама і праз распрацоўку новых метадаў і алгарытмаў апрацоўкі і аналіза дадзеных. Распрацоўваюцца новыя метады больш эфектыўнай нармалізацыі словаў, транскрыбацыі і больш дакладнага вызначэння націскнога складу.

Вынікі даследванняў, праведзеных у гэтых накірунках, адлюстраваныя ў работах М. Портэра (M. Porter), П. Вілета (P. Willett), Дж. Плісана (J. Plisson), К. Манінга, Н.В. Лукашевич, Ю. Ліўшыца, Г.В. Рымскага, А. Вакуліч-Дэя (A. Wakulicz-Deja), А. Даўні (A. Downey) і інш.

Другі раздзел прысвечаны аналізу існуючых мадэляў і метадаў пабудовы лінгвістычных інфармацыйна-пошукавых сістэм.

У задачы ІПС уваходзяць захоўванне вялікага аб’ёму дадзеных, а таксама арганізацыя доступу да гэтых дадзеных: даданне, рэдагаванне і выдаленне. Для гэтага ІПС выкарыстоўваюць базы дадзеных (БД). Разгледжаныя найбольш папулярныя мадэлі дадзеных: іерархічная, сеткавая і рэляцыйная. У выніку аналізу быў зроблены выбар на карысць рэляцыйнай мадэлі. Яна з’яўляецца простай і даступнай для разумення. Яна мае строгія правілы праектавання, якія базуюцца на матэматычным апараце, і поўную незалежнасць дадзеных. Прадметная галіна лінгвістычнай ІПС цалкам падыходзіць для фармалізацыі ў выглядзе табліц.

Іншай задачай ІПС з’яўляецца інфармацыйны пошук. Інфармацыйны пошук уключае ў сябе пошук дакументаў, апрацоўку вынікаў пошуку, а таксама пытанні мадэлявання, класіфікацыі і фільтрацыі дакументаў, праектаванне архітэктур пошукавых сістэм і карыстальніцкіх інтэрфейсаў. Для эфектыўнага пошуку і апрацоўкі вынікаў гэтага пошуку неабходна правільна выбраць пошукавую мадэль. Былі разгледжаныя булева, вектарная і імавернасная мадэлі інфармацыйнага пошуку. У выніку аналізу было вырашана выкарыстоўваць вектарную мадэль. Яна дазваляе рэалізаваць упарадкаваную выдачу вынікаў,

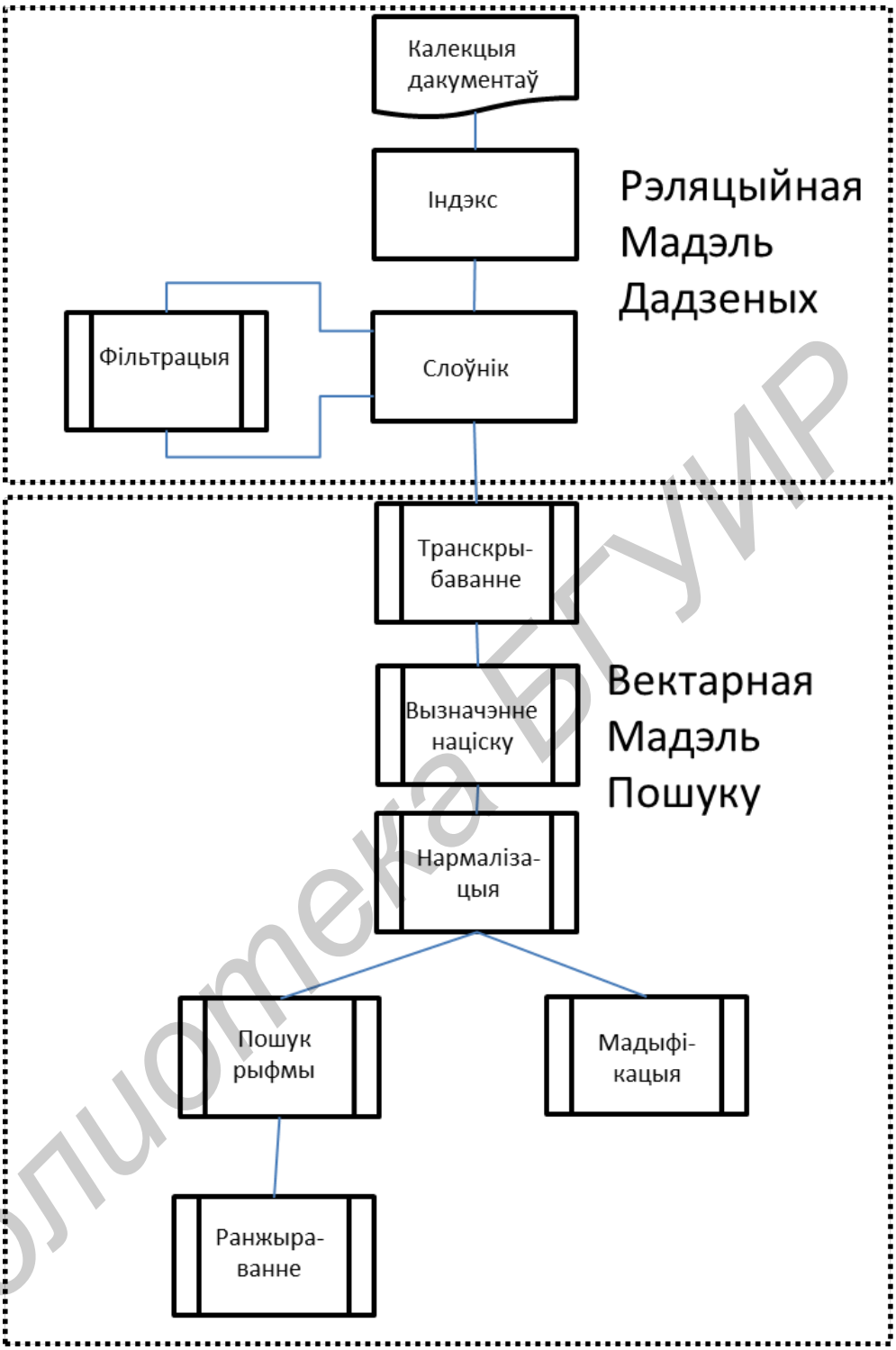
пры гэтым спосаб вылічэння рэлевантнасці лёгка змяняецца ў залежнасці ад пастаўленай задачы. Вага ўмоў запыту таксама можа адрозніваца. Для вызначэння рэлевантнасці паміж вектарамі запыта і дакумента выкарыстоўваецца гэтак званая косінусная мера:

$Sim(q,d_i,) = (\sum w_{qt*} w_{dt}) /(\sqrt{(\sum^{w2qt} * \sum^{w2dt})}$

Пошук павінны выконвацца не толькі па дакладным супадзенні, а і па ўсіх магчымых словаформах. Адпаведна, паўстае патрэба нармалізацыі словаў запыту і тэрмаў у слоўніку. Стэмінг (нармалізацыя) – гэта працэс знаходжання асновы слова для зададзенага зыходнага слова. Аснова неабавязкова супадае з марфалагічным коранем слова. Былі разгледжаныя розныя метады стэмінгу: пошук па табліцы, стэмінг Портэра, афікс-стэмеры з падстаноўкай, Ripple-down rules. У выніку іх аналізу была зробленая выснова аб тым, што ні адзін з метадаў не з'яўляецца па-асобку ўніверсальным рашэннем, і таму было прынята рашэнне выкарыстоўваць іх усе ў адным гібрыдным метадзе, якія спалучыць іх перавагі і пазбегне недахопаў.

Яшчэ адна задача лінгвістычных ІПС – неабходнасць аўтаматычнага вызначэння націску ў словах. Хоць сістэма і базуецца на выкарыстанні слоўніка, у якім усе словы ўжо маюць вызначаны націск, яна павінна ўмець працаваць і з незнаёмымі словамі. Сістэма павінна знаходзіць націскі ў такіх словах з высокай ступенню імавернасці. Былі прааналізаваныя некаторыя падыходы, якія выкарыстоўваюцца ў алгарытмах вызначэння націску: выкарыстанне формул, нармалізацыя, прадказанне на аснове статыстыкі. У выніку праведзенага аналізу было вырашана гэтаксама выкарыстоўваць усе падыходы ў межах вялікага алгарытма, які будзе паслядоўна-рэкурсіўна прымяняць усе апісаныя метады для дасягнення станоўчага выніку.

Трэці раздзел прысвечаны пабудове лінгвістычнай інфармацыйна-пошукавай сістэмы. На малюнку 1 адлюстраваная агульная схема архітэктуры ствараемай лінгістычнай ІПС. Сістэма складаецца з наступных кампанентаў: калекцыі дакументаў, індэкса, слоўніка, модуля транскрыпцыі, модуля вызначэння націску, модуля нармалізацыі, модуля ранжыравання вынікаў, аперацый мадыфікацыі калекцыі дакументаў і пошуку рыфмы.



Малюнак 1. Архітэктура лінгвістычнай ІПС.

У раздзеле апісаны падыход да эфектыўнага захоўвання вялікіх аб’ёмаў інфармацыі праз індэксацыю. Таксама прапанаваныя падыходы для хуткага пошуку інфармацыі. Распрацаваны метад пошуку рыфмы, які ўключае ў сябе нармалізацыю (стэмінг), Вызначэнне націску, транскрыбаванне, фільтрацыя (булевы пошук па базе дадзеных усіх патэнцыйных варыянтаў рыфмы) і ранжыраванне (вектарны пошук з ранжыраваннем сярод знойдзеных на папярэднім этапе для выдачы карыстальніку найбольш рэлевантных вынікаў).

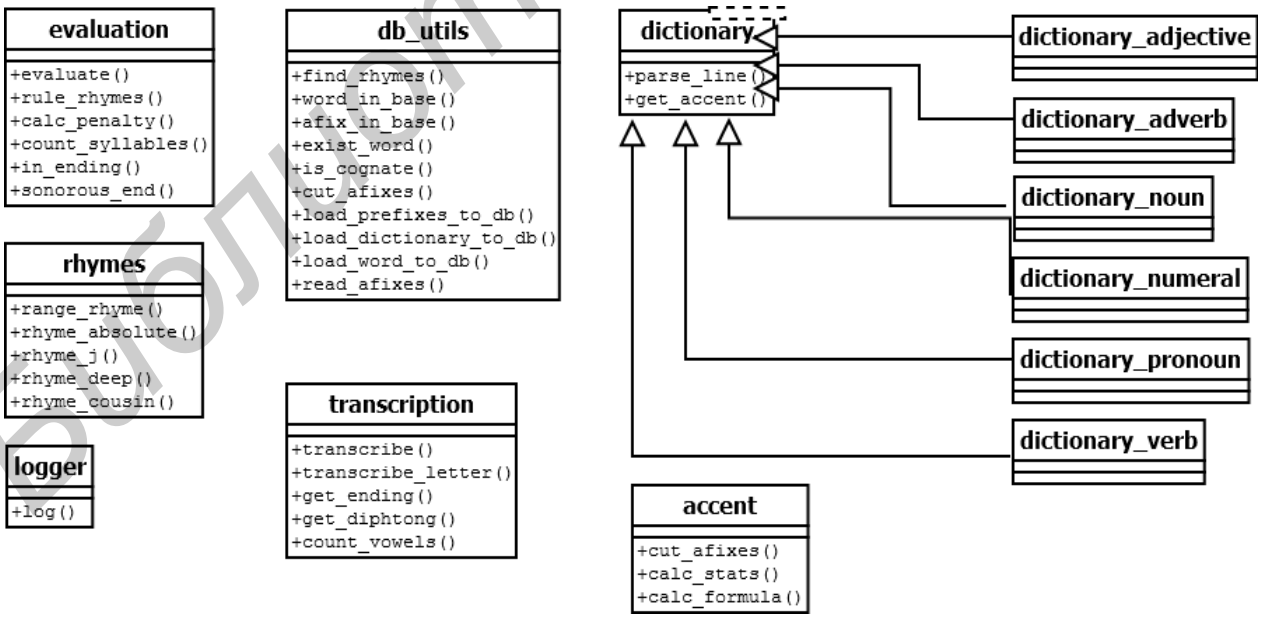
Стэмінг уключае ў сябе шэраг падыходаў, сярод якіх – мадыфікацыі стэмінгу Портэра, афікс-стэмеры з падстаноўкай і Ripple-down rules [1].

Таксама прапанаваны гібрыдны алгарытм пошуку націскнога склада. Схема алгарытма прыведзеная на малюнку 2:

Малюнак 2. Алгарытм пошуку націскнога склада.

Алгарытм складаецца з марфалагічнага аналіза, вызначэння стандартных афіксаў, прадказання на аснове статыстыкі і прадказання на аснове формул [2].

У **чацвёртым раздзеле** разгледжаная практычная рэалізацыя праграмнага сродку лінгвістычнай інфармацыйна-пошукавай сістэмы. Дыяграма класаў праграмнага сродку прыведзеная на малюнку 3.



Малюнак 3. Дыяграма класаў праграмнага сродку

Атрыманы праграмны сродак па шэрагу параметраў пераўзыходзіць аналагі. Па-першае, сістэм, якія працуюць з беларускай мовай, у адкрытым доступе проста не існуе. Па-другое, быў распрацаваны метад вызначэння

націску ў словах, заснаваны на шэрагу розных падыходаў, які дазваляе з вельмі высокай імавернасцю вызначаць націскны склад, што забяспечвае пабудовы правільнай транскрыпцыі. У большасці існуючых ІПС па пошуку рыфмы улік націску ці адсутнічае наогул, ці націск прастаўляецца з нізкай дакладнасцю. Патрэцяе, быў напісаны эфектыўны рэкурсіўны метад стэмінгу, які дазваляе вызначаць аднакаранёвыя рыфмы і пры гэтым не рабіць памылак залішняга адсячэння. Ні адна з ІПС па пошуку рыфмы ў вольным доступе на рускай мове не ўмее адрозніваць аднакаранёвыя рыфмы. Была створаная сістэма ранжыравання і класіфікацыі вынікаў пошуку, якая дазваляе групаваць рыфмы па тыпах і выдаваць уверсе пошукавай выдачы найбольш моцныя і дакладныя рыфмы. Існуючыя пошукі ў асноўнай сваёй масе альбо не маюць ранжыравання наогул, альбо яно недастаткова празрыстае і паказвае невысокія вынікі.

ЗАКЛЮЧЭННЕ

Асноўныя навуковыя вынікі дысертацыі

1. Прапанаваная архітэктура лінгвістычнай інфармацыйна-пошукавай сістэмы для вырашэння задач аўтаматызацыі пошуку рыфмы з улікам рэальнага гучання слова. Пры гэтым праводзіцца нармалізацыя слова, вызначэнне націскнога складу і пошук аднакаранёвых рыфм.
2. Распрацаваны метад нармалізацыі слова, які карэктна вызначае пачатковую форму слова і робіць марфалагічны аналіз для далейшага пошуку рыфмы.
3. Распрацаваны алгарытм пошуку націскнога складу ў незнаёмых сістэме словах, які ўлічвае вынікі марфалагічнага аналізу, выкарыстоўвае дадзеныя па стандартных афіксах, вынікі прадказанняў на аснове статыстыкі і на аснове формул. У выніку алгарытм паказвае вельмі высокую эфектыўнасць.
4. Распрацаваны праграмны сродак Лінгвістычная інфармацыйна-пошукавая сістэма “Вершнік” для пошуку рыфмы. Праграмны сродак мае форму вэб-сэрвіса і просты карыстальніцкі інтэрфейс.

Рэкамендацыі па практычным выкарыстанні вынікаў

1. Атрыманыя вынікі фарміруюць тэарэтычную і практычную базу для распрацоўкі ПЗ камп’ютарных сістэм для рашэння задач пабудовы інфармацыйна-пошукавых сістэм. Яны могуць быць скарыстаныя для мадыфікацыі і паляпшэння існуючых сістэм, а таксама для распрацоўкі новых.

Распрацаваныя метады і алгарытмы пошуку націскнога складу, нармалізацыі, транскрыбацыі могуць прымяняцца не толькі ў інфармацыйна-пошукавых сістэмах для вершаскладання, але і ў электронных перакладчыках, сістэмах сінтэзу маўлення і аналізу галасавых каманд і ў іншых сістэмах, звязаных з аналізам жывой чалавечай мовы.

СПІС АПУБЛІКАВАНЫХ РАБОТ

1. Філіпчык, А.В. Алгарытм транскрыбавання лексічных адзінак у лінгвістычнай інфармацыйна-пошукавай сістэме / А.В. Філіпчык // Материалы IV Республиканской научной конференции студентов, магистрантов и аспирантов, ГГУ им. Ф.Скорины, 15 апреля 2015г. – Гомель, 2015. – ч.3. – с. 190-192.

2. Філіпчык, А.В. Алгарытм аўтаматычнага вызначэння месца націску ў невядомых словах у лінгвістычнай інфармацыйна-пошукавай сістэме / А.В. Філіпчык // Материалы 51-й научной конференции аспирантов, магистрантов и студентов, БГУИР, 13-17 апреля 2015г. – Минск, 2015. – с. 77-78.

3. Филипчик, А.В. Стемминг слов в лингвистической информационно-поисковой системе / А.В. Филипчик // Апробация.– 2016. №2 (в печати).