OSTIS-2014
(Open Semantic Technologies for Intelligent Systems)

УДК 004.822:514

КАМПАНЕНТ ДЛЯ АНАЛІЗУ ЧАСЦІН МОВЫ БЕЛАРУСКІХ СЛОЎ ВА ЎМОВАХ АБМЕЖАВАНЫХ СІСТЕМНЫХ РЭСУРСАЎ

Шчурко М.Л.[*], Гецэвіч Ю.С.[**], Пакладок Д.А.[**]

[*] *Беларускі дзяржаўны ўніверсітэт інфарматыкі і радыёэлектронікі, Мінск, Рэспубліка Беларусь*

maxe1.first@gmail.com

[**] *Аб'яднаны інстытут праблем інфарматыкі НАН Беларусі, Мінск, Рэспубліка Беларусь*

yury.hetsevich@gmail.com

dima.pokladok@gmail.com

Апісваецца стварэнне кампанента для аўтаматычнага вызначэння часціны мовы слова па фіксаванай даўжыні заканчэння слова для 3, 4, 5 сімвалаў. Вынікі тэставанняў паказалі максімальную дакладнасць 65,3% пры фіксаванай даўжыні 5 сімвалаў. Прыведзена паслядоўнасць крокаў для ўжывання гэтага алгарытму на платформе J2ME, якая з'яўляецца тыповай платформай з абмежаванымі сістэмнымі рэсурсамі.
Ключавыя словы: кампанент; сінтэз маўлення па тэксце; мабільны сінтэзатар маўлення па тэксце.

Уводзіны

У цяперашні час немагчыма ўявіць жыццё і працу без мабільных прылад, такіх як тэлефоны, планшэты, інтэрактыўныя тэле-, аўдыясістэмы і розныя сістэмы кіравання тэхнікай. Дзякуючы вялікай колькасці задач, праграмнае забеспячэнне для іх актыўна развіваецца.

Падчас распрацоўкі праграмнага забеспячэння для гэтых прылад трэба памятаць, што яны маюць абмежаваны набор рэсурсаў памяці і камандаў, таму стандартныя і відавочныя рашэнні не заўсёды падыходзяць да гэтых платформ.

Адной з заўсёды запатрабаваных задач з'яўляецца сінтэз маўлення. А адной з падзадач у ім з'яўляецца вызначэнне часцін моў слоў сказу. Гэта нескладана зрабіць, калі мець поўны слоўнік слоў з пазнакамі часцін мовы, але для прылад з абмежаванай колькасцю памяці гэты спосаб нязручны, таму што гэтыя слоўнікі маюць вялікі памер [Цирульник, 2012]. Відавочна, што калі нельга выкарыстоўваць нейкі агульны слоўнік, то трэба ўжываць алгарытм, які зможа з высокай верагоднасцю вызначыць часціну мовы ўсіх слоў у тэксце.

Прыкладам такога алгарытму з'яўляецца вынаходжанне часціны мовы з дапамогай базы пар "заканчэнне – часціна мовы". Сутнасць алгарытма заключаецца ў вызначэнні часціны мовы слова па яго заканчэнню. Пад заканчэннем тут трэба разумець нейкую фіксаваную колькасць літар ад канца слова (ці ўсё слова, калі яго даўжыня менш за гэту колькасць).

У гэтым артыкуле даследаваная адпаведнасць гэтага алгарытму для беларускай мовы і прапанаваная рэалізацыя для платформы J2ME [Oracle, 2013] у выглядзе модуля для сістэмы сінтэзу маўлення па тэксце для мабільных тэлефонаў [Цирульник, 2012]. Гэта платформа выкарыстоўваецца зараз не вельмі шырока, але абмежаванасць тэлефонаў у сістэмных рэсурсах, якія яе ўжываюць, гарантуе, што алгарытм зможа працаваць на сучасных платорфмах, такіх як Google Android, Apple iOS, BlackBerry OS, Windows Phone і інш, а таксама проста ўбудоўвацца як кампанент у іншыя інтэлектуальныя сістэмы.

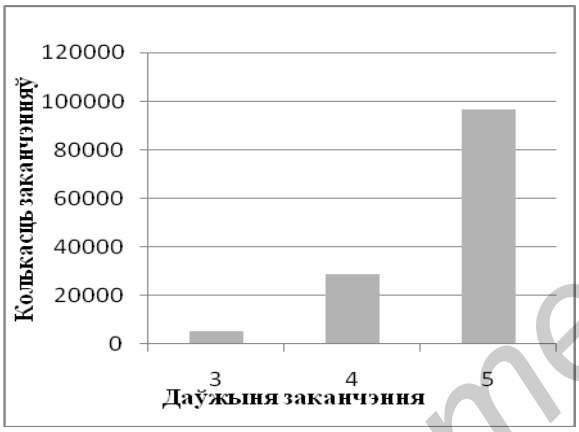
Праблема вызначэння часціны мовы слова не з'яўляецца новай: існуюць і іншыя алгарытмы. Частка з іх ужывае маркаўскія ланцугі [Paul Taylor, 2009], частка – нейронныя сеткі [Nakagawa, Kudoh, Matsumoto, 2001]. Але складанасць сучасных алгарытмаў можа стаць істотным абмежаваннем ці наогул спыніць працу сінтэзу маўлення на прыладзе, якая абмежаваная ў рэсурсах (вызначэнне часціны мовы – толькі адна з аперацый патрэбных для сінтэзу).

1. Даследаванне ідэнтыфікацыі часціны мовы слова па заканчэнні слова

Асноўным рэсурсам для алгарытма з'яўляецца слоўнік, у якім для кожнага слова была вызначана часціна мовы [Hetsevich, 2012]. Усяго ў гэтым слоўніку 2118455 словаўжыванняў. На яго базе будуецца новы слоўнік S, які складаецца з пар "заканчэнне – часціна мовы".

Для змяншэння аб'ёму слоўніка S было праведзена адмысловае даследаванне. Яго сутнасць складаецца ў вызначэнні аптымальных суадносін паміж аб'ёмам слоўніка S і дакладнасцю для выкарыстання слоўніка S у алгарытме вызначэння часцін мовы. У даследаванні разгледжана некалькі слоўнікаў S з заканчэннямі рознай даўжыні, а менавіта: не больш за 3, не больш за 4 і не больш за 5 літар. Яно складалася з трох этапаў:

1) Са слоўніка былі выбраныя ўсе магчымыя заканчэнні рознай даўжыні. Колькасць атрыманых заканчэнняў пэўнай даўжыні адлюстраваная на малюнку 1.



Малюнак 1 – Колькасць заканчэнняў пэўнай даўжыні

2) Для кожнага заканчэння пэўнай даўжыні былі ўзятыя адпаведныя словы-прыклады, якія яго маюць, і часціны мовы, да якіх належаць гэтыя словы. У выніку былі атрыманы файлы наступнага выгляду для кожнай даўжыні заканчэнняў:

Заканчэнне1 (слова11 часціна мовы11) (слова12 часціна мовы12) ...
Заканчэнне2 (слова21 часціна мовы21)

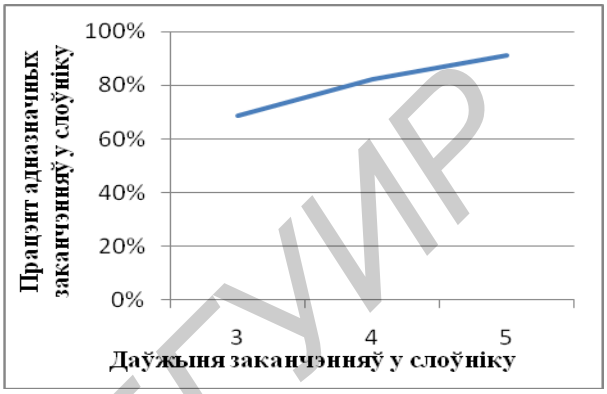
дзе першы індэкс гэта нумар радка, другі – нумар пары "слова – часціна мовы" ў гэтым радку.

3) У файлах, атрыманых на кроку 2, былі вызначаныя два віды заканчэнняў – адназначныя і шматзначныя.

Адназначнае заканчэнне – заканчэнне, якое характэрна толькі для адной часціны мовы. Напрыклад, заканчэнне *„-фатар'*, адназначна вызначае часціну мовы *назоўнік* (трыўмфатар). Шматзначныя заканчэнні, характэрныя для некалькіх часцін мовы. Напрыклад, заканчэнне *„–чнага'* можа сустракацца ў лічэбнікаў (*шасцітысячнага*) і назоўнікаў (*падапечнага*).

Праз шматзначнае заканчэнне немагчыма вызначыць часціну мовы адназначна без дадатковых рэсурсаў і алгарытмаў. Таму ўсе шматзначныя заканчэнні былі выдаленыя. Атрыманыя файлы з парамі "заканчэнне – часціна мовы" з'яўляюцца слоўнікамі, якія падыходзяць для адназначнага вызначэння часціны мовы слова ці для канстатацыі, што часціну мовы вызначыць нельга.

На малюнку 2 паказана залежнасць колькасці адназначных заканчэнняў у слоўніку ад іх даўжыні.



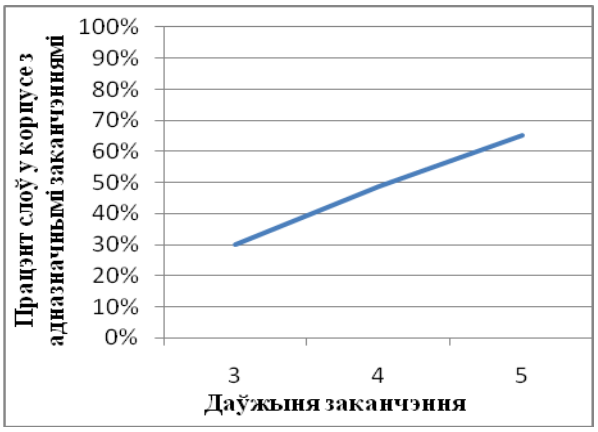
Малюнак 2 – Залежнасць колькасці адназначных заканчэнняў у слоўніку ад іх даўжыні

З малюнка 2 бачна, што пры павелічэнні даўжыні заканчэнняў слоў павялічваецца колькасць адназначных часцін мовы, якія могуць быць вызначаныя па заканчэннях. Гэта значыць, павялічваецца дакладнасць слоўніка, але разам з гэтым павялічваецца і яго аб'ём.

Такім чынам, на базе даследавання былі атрыманыя слоўнікі для трох-, чатырох- і пяцілітарных заканчэнняў, па якіх можна адназначна вызначыць часціну мовы.

2. Тэставанне дакладнасці слоўнікаў "заканчэнне слова – часціна мовы" на корпусе тэкстаў

Для тэставання быў узяты размечаны часцінамі мовы корпус памерам 284774 слоў з тэкстаў розных стыляў: навуковага, мастацкага, афіцыйна-дзелавога.



Малюнак 3 – Працэнт слоў у корпусе з адназначнымі заканчэннямі

На малюнку 3 адлюстраваны працэнт адназначных заканчэнняў у корпусе тэкстаў у залежнасці ад даўжыні гэтых заканчэнняў. Па іх можна меркаваць пра колькасныя паказчыкі магчымасцяў вызначаць часціны мовы з дапамогай распрацаваных слоўнікаў.

У табліцы 1 сабраныя ўсе лічбы і вынікі тэставанняў для аднолькавых тэкстаў з прымяненнем розным па напаўненні слоўнікаў “заканчэнне – часціна мовы”.

Табліца 1 – Вынікі тэставанняў на размечаным корпусе тэкстаў

Параметры тэставанняў	Тэст 1	Тэст 2	Тэст 3
Даўжыня заканчэнняў	3	4	5
Усяго слоў у корпусе тэкстаў	284774	284774	284774
Слоў з адназначнымі заканчэннямі	85076	138364	186035
Адносная колькасць слоў з адназначнымі заканчэннямі	29,9%	48,6%	65,3%

Вынікі тэставанняў паказалі максімальную дакладнасць 65,3%. Яе “неадпаведнасць 100% дакладнасці” абумоўлена папярэднім рашэннем адносна пастаўленай задачы: былі адкінутыя шматзначныя заканчэнні, што ўплывае на адкіданне шматзначных слоў. Вынікі тэставанняў паказалі паступовае памяншэнне дакладнасці да 29,9%, бо дакладнасць змяншаецца пры выкарыстанні слоўнікаў зменшаных даўжынь заканчэнняў у слоўніках ад 5 да 3.

3. Выкарыстанне слоўнікаў пад J2ME

Нягледзячы на прастату тэарытычнага і эксперыментальнага выкарыстання (для стацыянарнага запуску праграмы на J2ME) пабудаваных слоўнікаў для вырашэння пастаўленай задачы вызначэння часцін мовы слоў застаецца праблема выкарыстання слоўнікаў у рэальных мабільных прыстасаваннях. Пажадана, каб аб'ём слоўніка быў меншы чым 900 кБ. Таму патрэбны нейкі дадатковы алгарытм, каб паменшыць колькасць месца, якое займае масіў дадзеных слоўніка (аднаго з трох распрацаваных). Ідэя наступнага алгарытму ў тым, каб прадставіць кожны сімвал заканчэння і адпаведнага заканчэнню кода не двума байтамі, як гэта робіцца ў Java, а нейкай мінімальнай колькасцю біт фіксаванай даўжыні.

Крокі алгарытму:

1) Выбраць слоўнік з фіксаванай даўжынёй заканчэння F (паводле другога раздзелу артыкула).

2) Стварыць алфавіт з розных сімвалаў, якія ўжываюцца ў заканчэннях і пранумараваць сімвалы, пачынаючы з нуля.

3) Пранумараваць розныя часціны мовы слоўніка, пачынаючы з нуля.

4) Вызначыць мінімальную колькасць біт, якая патрэбная, каб пазначыць нумар любога сімвалу альфабэту $N_{альф}$. Напрыклад, для таго каб абазначыць альфабэт з 29 сімвалаў, дастаткова 5 біт ($2^4 < 28 < 2^5$; 28 таму што адлік пачынаецца з нуля).

5) Аналагічна кроку 4 вызначыць такую колькасць біт $N_{часц}$, якая дазволіць вызначыць нумар любой часціны мовы.

6) Каб усе заканчэнні былі аднолькавай даўжыні, дапоўніць іх прабеламі, дзе патрэбна.

7) Закадаваць усе пары “заканчэнне – часціна мовы” паслядоўна ў масіў байтаў. Заканчэнне павінна адпавядаць фармату:
нумар_літары1нумар_літары2...нумар_літарыN

Выніковая паслядоўнасць павінна адпавядаць фармату:
заканчэнне1нумар_часціны_мовы1заканчэнне2нумар_часціны_мовы2...заканчэннеNнумар_часціны_мовыN.

Даўжыня аднаго заканчэння будзе $N_{альф} \times$ даўжыня_канчатку + $N_{часц}$ біт замест стандартных $8 \times 2 \times$ даўжыня канчатку + 8×1 біт, калі захоўваць заканчэнні ў радковым выглядзе. Дзякуючы таму, што $N_{альф}$ будзе меншым за 8×2 (для беларускай мовы $N_{альф} = 6$) і $N_{часц}$ таксама будзе менш за 8 (для 10 часцін мовы $N_{часц} = 4$), а даўжыня канчатку мінімум 3, то на выхадзе атрымаецца слоўнік SS меншых памераў, чым слоўнік да закадавання ўсіх пар “заканчэнне – часціна мовы”.

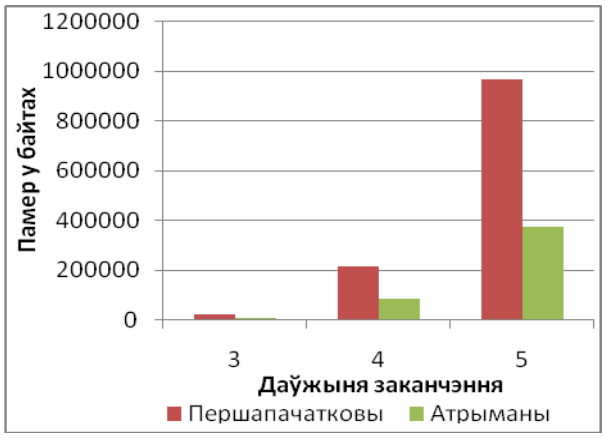
8) Модуль у тэлефоне прымае як дапаможны рэсурс слоўнік SS. З кожнага ўваходнага слова, для якога патрэбна ідэнтыфікаваць часціну мовы, выдзяляецца заканчэнне адпаведнай даўжыні F. (Калі не хапае літар для выдзялення заканчэння, то астатнія пазіцыі запаўняюцца прабеламі.) Далей гэтае заканчэнне правяраецца на наяўнасць па слоўніку SS, калі такое èсць, то выдаецца за адказ адпаведная часціна мовы, калі такога няма, то выдаецца адказ UNKNOWN.

Параўнанне памераў першапачатковых (паводле раздзелу 2) і атрыманых слоўнікаў SS (толькі файлаў пар “заканчэнне – часціна мовы” без памеру альфабэту і іншай невялікай службовай інфармацыі) прыводзіцца на малюнку 4. Лічбы для гэтага малюнку прыведзены ў табліцы 2. На малюнку бачна, што алгарытм, які выкарыстоўваецца, значна змяншае памер слоўнікаў для любой фіксаванай даўжыні заканчэння.

Табліца 2 – Зыходныя і выніковыя памеры слоўнікаў

Даўжыня заканчэння	Зыходны памер слоўніка, байт	Выніковы памер слоўніка, байт	Размер зменшаны на, %
3	24710	9709	60,71
4	213156	82894	61,11
5	966614	373465	61,36

Найбольш прымальнай даўжынѐй канчатку для мабільнай платформы з‟яўляецца даўжыня не большая за 4 літары з-за аптымальных суадносін “аб'ѐм слоўніку – дакладнасць вырашэння пастаўленай задачы”.



Малюнак 4 – Параўнанне памеру атрыманага і першапачатковага масіву дадзеных

Заключэнне

Такім чынам, была пастаўлена і вырашана з пэўнай дакладнасцю задача прадказання часціны мовы слоў на прыладзе з абмежаванымі рэсурсамі. Для гэтага быў прапанаваны алгарытм ідэнтыфікацыі часціны мовы па заканчэннях слоў вядомага электроннага граматычнага слоўніка. У залежнасці ад выкарыстання слоўніка для фіксаванай даўжыні заканчэння слова ад 3 да 5 літар дакладнасць працы алгарытма мяняецца ад амаль 30% да 65%. Для выкарыстання ў рэальных мабільных прыстасаваннях быў прапанаваны якасны алгарытм скарачэння выкарыстання памяці для захавання слоўнікаў у залежнасці ад фіксаванай даўжыні заканчэння слова для не больш за 3, 4, 5 літар. Алгарытм кампрэсіі даў магчымасць скараціць аб‟ѐм слоўніка на 60%, 61%, 61% адносна першапачаткова атрыманых.

Для павышэння дакладнасці працы алгарытма па вызначэнні часціны мовы слоў трэба ўжываць у слоўніку таксама шматзначныя заканчэнні. Аўтарамі прапануецца ў будучым выбіраць сярод часцін мовы, характэрных для шматзначных заканчэнняў, часціну мовы, якая найбольш часта ўжываецца з гэтым заканчэннем у тэкстах.

Спіс літаратуры

[Цирульник, 2012] Цирульник, Л.И. Алгоритмы создания и пополнения грамматического словаря русского языка для синтеза речи по тексту / Л.И. Цирульник, В.В. Веремей // Информатика. – 2012. – № 1 (31). – С. 27–38.

[Oracle, 2013]Java ME Landing Page // Oracle [Electronic recourse]. – 2013. – Mode of access: http://www.oracle.com/technetwork/java/javame/index.html. – Date of access: 23.07.13.

[Цирульник, 2012] Цирульник, Л.И. Система синтеза речи по тексту для мобильных телефонов / Л.И. Цирульник, Д.А. Покладок // Речевые технологии. – 2010. – № 1. – С. 81-90.

[Paul Taylor, 2009] Paul Taylor Text-to-Speech Synthesis // Paul Taylor // Cambridge University Press – 2009. – P. 89 – 93.

[Nakagawa, Kudoh, Matsumoto, 2001] Tetsuji Nakagawa, Taku Kudoh and Yuji Matsumoto Unknown Word Guessing and Part-of-Speech Tagging Using Support Vector Machines // Tetsuji Nakagawa, Taku Kudoh and Yuji Matsumoto // Graduate School of Information Science, Nara Institute of Science and Technology. - 2001

[Hetsevich, 2012] Hetsevich, Y. Overview of Belarusian And Russian dictionaries and their adaptation for NooJ / Y. Hetsevich, S. Hetsevich // Automatic Processing of Various Levels of Linguistic Phenomena: Selected Papers from the NooJ 2011 Intern. Conf. / eds. Vučković Kristina, Bekavac Božo, Silberztein Max. – Newcastle : Cambridge Scholars Publishing, 2012. – P. 29 – 40.

COMPONENT FOR PART-OF-SPEECH TAGGING OF BELARUSIAN WORDS WITH LOW SYSTEM REQUIREMENTS

Shchurko M.L. *, Hetsevich Y.S. **,
Pakladok D.A. **

**Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus*
maxe1.first@gmail.com

***United Institute of Informatics Problems, National Academy of Sciences, Minsk, Republic of Belarus*
yury.hetsevich@gmail.com
dima.pokladok@gmail.com

Introduction

This article describes the problem of part-of-speech tagging on devices with architecture limitations of memory and processor power like mobile phones, tablet PC, etc.

The problem is that system requirements for modern algorithms don‟t fit these devices.

Main Part

In this article was researched algorithm of guessing part-of-speech by the ending of a word for endings that have unambiguous corresponded part-of-speech. Also was proposed algorithm of compression pairs “ending – part-of-speech”.

Mentioned algorithms were tested with different ending size (from 3 to 5). On tested text it has 30%, 45%, 65% accuracy for 3, 4, 5 ending size accordingly. It decreases size of pairs base by 60%, 61%, 61% for 3, 4, 5 ending size accordingly.

Conclusion

Researched algorythm is not accurate enough. Should be proposed another algorithm that statistically chooses part-of-speech for endings which have polysemantic corresponded parts-of-speech. Compression algorithm allows to store greater “ending – part-of-speech” base on a device.