Объединенный институт проблем информатики
Национальной академии наук Беларуси

XXI Международная
научно-техническая конференция

РАЗВИТИЕ ИНФОРМАТИЗАЦИИ И ГОСУДАРСТВЕННОЙ СИСТЕМЫ НАУЧНО-ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ

РИНТИ-2022

17 ноября 2022 г., Минск

Доклады

Минск
ОИПИ НАН Беларуси
2022

УДК 002; 004

Развитие информатизации и государственной системы научно-технической информации (РИНТИ-2022) : доклады XXI Международной научно-технической конференции, Минск, 17 ноября 2022 г. – Минск : ОИПИ НАН Беларуси, 2022. – 408 с. – ISBN 978-985-7198-12-2.

Представлены доклады XXI Международной научно-технической конференции «Развитие информатизации и государственной системы научно-технической информации» (РИНТИ-2022), Минск, 17 ноября 2022 г., в которых рассмотрены вопросы и перспективы формирования единого цифрового пространства научной отрасли, основные результаты научно-методического обеспечения развития информатизации в НАН Беларуси, проблемы и пути решения государственного суверенитета в цифровую эпоху, модернизация содержания и цифровая трансформация университетского ИТ-образования, сценарии развития цифровой экосистемы земельного администрирования в Беларуси, перспективы интеллектуальной собственности, передовых технологий и искусственного интеллекта в республике, страницы истории белорусской вычислительной техники и др.

Рассмотрены вопросы научно-методического, информационного, технологического и правового обеспечения цифровой трансформации, проектирования и внедрения автоматизированных систем научно-технической информации, библиотечно-информационных систем и технологий, публикационной активности ученых, а также направления развития искусственного интеллекта и когнитивных технологий в информатизации.

Материалы конференции будут полезны специалистам в области информационно-коммуникационных технологий, занимающимся научно-методическим обеспечением информатизации и решением задач построения ИТ-страны, цифровой экономикой, разработкой и внедрением автоматизированных информационных систем управления, систем научно-технической информации, автоматизированных библиотечно-информационных систем и технологий, а также развитием информационной инфраструктуры Беларуси и других стран, реализацией проектов государственных и отраслевых программ в сфере информатизации.

Одобрены программным комитетом и печатаются по решению редакционной коллегии Объединенного института проблем информатики Национальной академии наук Беларуси.

Научные редакторы:

доктор военных наук, кандидат технических наук, доцент С. В. Кругликов,
кандидат технических наук, доцент Р. Б. Григянец,
кандидат технических наук, доцент В. Н. Венгеров

ISBN 978-985-7198-12-2

УДК 004.934

АЎТАМАТЫЧНАЕ ПЕРАЎТВАРЭННЕ БЕЛАРУСКАГА МАЎЛЕННЯ Ў ТЭКСТ

А. С. Трафімаў, Ю. С. Гецэвіч
Аб’яднаны інстытут праблем інфарматыкі НАН Беларусі, Мінск

Апісана распрацоўка сістэмы распазнавання адвольнага беларускага маўлення з выкарыстаннем сучасных архітэктур глыбокага навучання. Выбар мадэляў абумоўлены іх добрыми вынікамі па распазнаванні маўлення, зменшанымі патрабаваннямі да памераў анатаваных датасэтаў у параўнанні з папярэднімі мадэлямі, а таксама адсутнасцю вялікіх анатаваных корпусаў агучаных тэкстаў для беларускай мовы.

Уводзіны

Пераўтварэнне маўлення ў тэкст або распазнаванне маўлення (Speech-to-Text, STT або Automatic Speech Recognition, ASR (АРМ)) – гэта адна з асноўных задач апрацоўкі натуральнага маўлення. Яе мэтай з’яўляецца аўтаматычная будова тэкставай транскрыпцыі для аўдыяфайла з запісаным маўленнем аднаго ці некалькіх людзей. Актуальнасць распрацоўкі беларускіх АРМ тлумачыцца адсутнасцю для беларускай мовы якасных сістэм распазнавання адвольнага маўлення. Атрыманне такой сістэмы дазволіць перайсці да распрацоўкі больш складаных моўных тэхналогій і пачаць інтэграваць іх у штодзённае жыццё. Прыкладам выкарыстання такой тэхналогіі з’яўляецца галасавы набор тэкста або стварэнне галасавых інтэрфэйсаў для кіравання прыладамі, у тым ліку стварэнне галасавых дапаможнікаў.

Эксперыменты па распазнаванні беларускага маўлення праводзіліся і раней. Так, супрацоўнікамі лабараторыі распазнавання і сінтэзу маўлення АІПІ НАН Беларусі праводзіліся эксперыменты па распазнаванні маўлення з абмежаваным слоўнікам [1, 2]. Таксама праводзіўся эксперымент па распазнаванні адвольнага маўлення і зборы патрэбнага для гэтай задачы датасэта [3], аднак у гэтым даследаванні прысутнічалі недахопы, якія перашкаджаюць будове ўстойлівай мадэлі і робяць немагчымым ацэнку яе якасці для выкарыстання ў штодзённым жыцці (напрыклад, для распазнавання маўлення, запісанага на вуліцы на мікрафон тэлефона). Менавіта ў эксперыменце [3] прысутнічалі наступныя недахопы: адсутнічалі формулы для метрык, выкарыстаных для ацэнкі якасці мадэляў; супярэчлівыя значэнні прыведзеных метрык; усе тэксты былі начытаны ў спецыяльна абсталяваным для запісу голаса пакоі; частка тэкстаў начытвалася прафесійнымі дыктарамі; моўная мадэль будавалася толькі на тэкстах, якія ўвайшлі ў датасэт (памер тэкставага корпуса быў невялікі); малы памер і варыятыўнасць датасэта, на якім ацэньвалася мадэль.

Такім чынам, задача распазнавання адвольнага беларускага маўлення не з’яўляецца вырашанай, таму патрабуецца пабудаваць якасную мадэль для распазнавання адвольнага (без абмежавання на распазнаваемыя словы) беларускага маўлення і сабраць неабходны для гэтага датасэт (корпус начытаных тэкстаў на беларускай мове). Датасэт (набор даных) мусіць складацца з гукавых файлаў з запісамі маўлення і адпаведных ім тэкставых транскрыпцый, а таксама мусіць быць дастатковага аб’ёму з як мага большай варыятыўнасцю дыктараў (пол, узрост, тэмп маўлення, артыкуляцыя, дыялекты, інш.) і ўмоваў запісу (наяўнасць фонавага шуму, рэха, недасканаласці мікрафонаў, інш.).

1. Збор датасэта для навучання мадэлі распазнавання маўлення

Каб сабраць датасэт патрэбнага памеру, было вырашана выкарыстаць краўдсорсынгавыя (ад англ. crowdsourcing) анлайн-платформы для калектыўнага агучвання падрыхтаванага загадзя корпуса тэкстаў. Такі падыход дазваляе кантраляваць працэс падрыхтоўкі тэкставага корпуса, прыцягваць да агучвання вялікую колькасць людзей і тым самым забяспечваць высокую варыятыўнасць галасоў, акцэнтаў, вымаўленняў ды акустычных умоваў запісу.

Для арганізацыі працэса збора даных была выкарыстана краўдсорсынгавая анлайн-платформа *Mozilla Common Voice* (https://commonvoice.mozilla.org/en/ datasets). Згодна ўмовам выкарыстання Common Voice, тэксты, запампаваныя на платформу, мусяць знаходзіцца ў публічнай прасторы (мець ліцэнзію CC-0). Таму пачатковы корпус тэкстаў на беларускай мове для запампоўвання на Common Voice быў створаны з тэкстаў артыкулаў беларускай Вікіпедыі. Пазней тэкставы корпус быў дапоўнены беларускімі мастацкімі творамі, напісанымі не пазней за 70 гадоў таму, што аўтаматычна пераводзіць іх у публічную прастору згодна з беларускім заканадаўствам.

Перад запампоўваннем на Common Voice тэксты апрацоўваліся і фільтраваліся. Былі абраны толькі сказы працягласцю не больш за 14 слоў, якія складаюцца з літар беларускага алфавіту (у тым ліку апострафа ‘), прагалаў, асноўных сімвалаў пунктуацыі (,.!?) і розных варыянтаў злучкоў (-, –, інш.), каб не ўскладняць дыктарам працэс начыткі. Выдаляліся сказы, яны змяшчалі такія словы, якія адсутнічаюць у граматычнай базе беларускай мовы і таксама сустракаюцца ў беларускай Вікіпедыі ≤ 60 разоў (каб улічыць адсутнасць у граматычнай базе некаторых распаўсюджаных у мове словаў і іх формаў). Дадаткова выдаляліся ўсе сказы з уласнымі імёнамі для палягчэння навучання мадэлі.

2. Аналіз сабранага датасэту

Сабраны на платформе Common Voice датасэт беларускай мовы версіі 8.0 (ад 19.01.2022) мае наступныя колькасныя прыкметы: агульная колькасць сабраных гадзін – 987, колькасць праверaных гадзін – 903, агульная колькасць дыктараў – 6160 чал., колькасць унікальных сказаў у датасэце – 347 010, колькасць праверaных аўдыязапісаў – 677 936, фармат захавання аўдыязапісаў – mp3, частасць дыскрэтызацыі – 32 кГц.

Перад публікацыяй датасэту на платформе Common Voice усе сабраныя даныя для кожнай мовы падзяляюцца на зафіксаваныя выбаркі. Сярод іх тры асноўныя: навучальная (train), валідацыйная (dev), тэставая (test). І чатыры дадатковыя: праверaныя (validated), памылковыя (invalidated), астатнія (other), адрынутыя (reported). Гэтыя выбаркі можна выкарыстоўваць у якасці бенчмарка для параўнання між сабою розных навучаных мадэляў.

Аднак для навучання найлепшай магчымай мадэлі распазнавання маўлення стандартная разбіўка даных не пасуе. Так, падчас працэса агучвання сказаў адзін і той жа сказ мог быць начытаны рознымі дыктарамі. Аднак у навучальную, валідацыйную, тэставую выбаркі файлы абіраюцца такім чынам, каб кожны сказ быў агучаны толькі адзін раз. Калі зняць гэта абмежаванне, то агульная колькасць аўдыяфайлаў у навучальнай, валідацыйнай і тэставай выбарках павялічыцца амаль удвая – з 345 909 да 677 936.

Каб ацаніць варыятыўнасць аўдыясігналаў, былі прааналізаваны аўдыязапісы з валідацыйнай выбаркі (з прычыны яе меншага памеру ў параўнанні з навучальнай выбаркай). Вынікі аналізу можна абагуліць і на ўсю навучальную выбарку, бо памеры

валідацыйнай і тэставай выбаркі з’яўляюцца статыстычна значнымі адносна памераў навучальнай выбаркі (99 % даверальны інтэрвал з хібай ≤ 1 %) (https://github.com/common-voice/CorporaCreator).

Кожны аўдыязапіс з валідацыйнай выбаркі апісваўся наступнымі прыкметамі:

– 95 % модуля значэнняў аўдыязапіса. Дазваляе ацаніць агульны ўзровень гучнасці аўдыязапіса і не ўлічваць пікі і выкіды (у адрозненне ад максімальнага па модулі значэння), якія могуць быць абумоўлены шумам, выбухнымі зычнымі (п, к, т) і г. д.;

– тэмпам маўлення, які падлічваўся як колькасць выбарак (фрэймаў) аўдыясігналу, падзеленую на колькасць літараў у сказе.

Далей для кожнага дыктара знаходзілася сярэдняе значэнне пабудаваных статыстык па агучаных ім сказах. Па атрыманых статыстыках для дыктараў была пабудавана дыяграма рассейвання, якая засведчыла шырокі спектр значэнняў статыстык для дыктараў, а значыць, і высокую варыятыўнасць сабраных гукавых даных.

3. Навучанне мадэлі распазнавання маўлення

Для распрацоўкі мадэлі распазнавання маўлення была абрана сучасная глыбокая нейрасеткавая архітэктура *wav2vec2* [4]. Яе асаблівасцю з’яўляецца пераднавучанне на корпусе неанатаваных даных (у рэжыме без настаўніка) для вывучэння спосабаў якаснага вылучэння прыкмет па ўваходным аўдыязапісе. Атрыманыя прыкметы выкарыстоўваюцца для далейшых падзадач, напрыклад для давучвання мадэлі пераводу маўлення ў тэкст. У якасці пераднавучанай мадэлі была абрана *facebook/wav2vec2-base* (https://huggingface.co/facebook/wav2vec2-base). Яе давучванне праводзілася на сабраным з дапамогай платформы Common Voice датасэце беларускага маўлення. Навучальная, валідацыйная і тэставыя выбаркі пакідаліся без змен: абмежаванне на колькасць агучванняў аднаго і таго ж сказа не здымалася, памер датасэта склаў 345 909 аўдыязапісаў.

Звычайна сістэмы распазнавання маўлення складаюцца з двух асноўных кампанентаў:

акустычнай мадэлі – уяўляе сабой блок сістэмы распазнавання маўлення, які па вылучаных з уваходнага аўдыясігналу прыкметах будуе паслядоўнасць фанем (або літар), вымаўленых з найбольшай імавернасцю;

моўнай мадэлі – патрэбна для пераводу атрыманага па ўваходным аўдыязапісе набора фанем або літар у набор найбольш імаверных слоў – выніковую транскрыпцыю.

3.1. Навучанне акустычнай мадэлі

Усе аўдыязапісы апрацоўваліся згодна наступнаму фармату: частасць дыскрэтызацыі 16 кГц, 1 канал (моназапісы). Тэкставыя транскрыпцыі прыводзіліся да ніжняга рэгістру; прыбіраліся ўсе сімвалы акрамя літар алфавіта і лічбаў; кожная паслядоўнасць сімвалаў-прагалаў (прагал, знак табуляцыі, інш.) замянялася на адзін прагал. У якасці функцыі стратаў для давучвання мадэлі на задачы распазнавання маўлення выкарыстоўвалася *Connectionist Temporal Classification, CTC* [5]. Аптымізацыя параметраў адбывалася з дапамогай алгарытма *AdamW* – выпраўленай версіі папулярнага алгарытма аптымізацыі *Adam*. Для аптымізацыі спажывання памяці выкарыстоўваўся метад *Gradient checkpointing*. Выбар найлепшай мадэлі праводзіўся з дапамогай метрыкі *Word Error Rate, WER* [6] на валідацыйнай выбарцы, далей найлепшая мадэль ацэньвалася на тэставай выбарцы.

Для праграмнай рэалізацыі навучання акустычнай мадэлі быў абраны папулярны фрэймворк для навучання NLP і ASR мадэляў *HuggingFace*, які з'яўляецца абгорткай над іншым фрэймворкам – *PyTorch*. Навучанне праводзілася на серверы з трыма відэакартамі *NVIDIA GeForce RTX 2080 Ti*. Памер батча для навучання і ацэнкі якасці мадэлі быў роўны 48 (16 элементаў у батчы на кожнай з трох відэакартаў). Сярэдні час на эпоху склаў каля васьмі гадзін. У сувязі з гэтым і з прымальнымі значэннямі метрык навучанне было спынена пасля пяці эпохаў. Ацэнка якасці (evaluation) і захаванне прамежкавых параметраў (checkpointing) праводзілася некалькі разоў на працягу кожнай эпохі.

Найлепшае значэнне WER на валідацыйнай выбарцы было дасягнута на апошнім чэкпоінце (0,176). Ён і быў абраны ў якасці найлепшай акустычнай мадэлі. Якасць абранай найлепшай акустычнай мадэлі была праверана на адкладзенай тэставай выбарцы. Значэнне Test WER склала 0,187.

3.2. Пабудова моўнай мадэлі

Для паляпшэння прадказанняў акустычнай мадэлі была пабудавана пяціграмная моўная мадэль з выкарыстаннем мадыфікаванага згладжвання Кнэсер – Нэй (modified Kneser – Ney smoothing). Для пабудовы такой мадэлі была выкарыстана папулярная бібліятэка *KenLM*. Моўная мадэль вучылася на корпусе тэкстаў з датасэта Common Voice 8.0. У корпус увайшлі ўнікальныя сказы з навучальнай выбаркі (train) і выбаркі validated без сказаў з dev, test выбарак, каб перадухіліць магчымыя выцёкі даных (data leakage). Агульная колькасць сказаў для пабудовы моўнай мадэлі склала 314 676.

Як згадвалася раней, важна, каб моўная мадэль будавалася на вялікім і разнастайным корпусе тэкстаў. 300 тыс. сказаў не з'яўляюцца шырокім тэкставым корпусам, здольным перадаць усе асаблівасці мовы і змясціць усе словаформы і іх ужыванні. Аднак збор патрэбнага большага па памеры тэкставага корпуса з'яўляецца асобнай аб'ёмнай і руплівай задачай. Таму ў межах дадзенай работы было вырашана пабудаваць моўную мадэль толькі на даступных для навучання сказах з датасэта Common Voice 8.0.

4. Ацэнка якасці выніковай сістэмы і стварэнне вэб-інтэрфэйса для дэманстрацыі работы мадэлі

У табліцы прыводзяцца значэнні метрыкі WER на валідацыйнай і тэставай выбарках разам з доляй цалкам распазнаных сказаў (транскрыпцый без памылак). Метрыкі прыводзяцца асобна для акустычнай мадэлі і для акустычнай мадэлі разам з моўнай мадэллю. Даданне моўнай мадэлі дазволіла зменшыць WER на тэставай выбарцы з 0,187 да 0,124. Канчатковы вынік – *test WER 0,124* (або 12,4 %) – з'яўляецца даволі добрым для мадэляў распазнавання.

Метрыкі выніковай мадэлі

Мадэль	dev WER	test WER	Доля распазнаных сказаў, test, %
Акустычная	0,1761	0,187	36,688
Акустычная + моўная	0,115	0,124	52,269

Пабудаваная акустычная мадэль разам з моўнай мадэллю былі загружаны на платформу *Hugging Face Hub* пад вольным доступам, што дазваляе любому зацікаўленаму выкарыстоўваць мадэлі далей (https://huggingface.co/ales/wav2vec2-cv-be). У дадатак на платформе Hugging Face Spaces быў рэалізаваны *вэб-інтэрфэйс*

(https://huggingface.co/spaces/ales/wav2vec2-cv-be-lm) для дэманстрацыі работы навучанай сістэмы распазнавання маўлення, якая складаецца з акустычнай і моўнай мадэляў. Інтэрфэйс дазваляе падаць альбо існуючы, альбо запісаны наўпрост у браўзеры аўдыязапіс на ўваход навучанай сістэме распазнавання і атрымаць пабудаваную транскрыпцыю.

Заключэнне

У выніку праведзенай працы была пабудавана новая для беларускай мовы сістэма распазнавання адвольнага маўлення высокай якасці (Test WER = 0,124 або 12,4 %), заснаваная на end-to-end архітэктуры з выкарыстаннем глыбокага навучання.

Для распрацоўкі дадзенай мадэлі APM быў сабраны вялікі корпус начытаных тэкстаў на беларускай мове з дапамогай платформы Mozilla Common Voice. Агульная працягласць сабраных аўдыязапісаў (на момант 19.01.2022) складае 987 гадзін (з іх 903 праверана), у агучванні якіх прынялі ўдзел 6160 дыктараў. Гэта першы з падобных датасэтаў такога памеру для беларускай мовы. Высокая варыятыўнасць сабраных даных як адносна дыктараў (пол, узрост, тэмп маўлення, іншыя асаблівасці вымаўлення), так і адносна ўмоваў запісаў (розныя мікрафоны, наяўнасць фонавага шуму, інш.) дазваляе навучыць сістэмы распазнавання маўлення працаваць ва ўмовах, набліжаных да тых, з якімі гэтым сістэмам давядзецца працаваць у штодзённым жыцці.

Праграмны код для навучання мадэляў (https://github.com/yks72p/stt_be) разам з пабудаванымі акустычнай і моўнай мадэлямі (https://huggingface.co/ales/ wav2vec2-cv-be) загружаны ў вольны доступ. Для дэманстрацыі работы пабудаванай сістэмы распазнавання маўлення быў створаны адмысловы вэб-інтэрфэйс на платформе Hugging Face (https://huggingface.co/spaces/ales/wav2vec2-cv-be-lm).

Спіс літаратуры

1. Гецэвіч, Ю. С. Распрацоўка кампанента распазнавання маўлення для натуральнага маўленчага інтэрфейсу / Ю. С. Гецэвіч, К. А. Нікалаенка, Л. І. Кайгародава // Открытые семантические технологии проектирования интеллектуальных систем = Open Semantic Technologies for Intelligent Systems (OSTIS-2015) : материалы V Междунар. науч.-техн. конф., Минск, 19–21 февр. 2015 г. – Минск : БГУИР, 2015. – С. 507–512.

2. Nikalaenka, K. Training algorithm for speaker-independent voice recognition systems using HTK / K. Nikalaenka, Y. Hetsevich // PRIP ‘2016: Pattern Recognition and Information Processing : Proc. of the 13th Intern. Conf., 3–5 Oct. 2016, Minsk, Belarus. – Minsk : Publishing Center of BSU, 2016. – P. 126–129.

3. Распрацоўка алгарытмаў дыктаранезалежнага распазнавання беларускага маўлення / Н. Д. Казлоўская [і інш.] // Развитие информатизации и государственной системы научно-технической информации (РИНТИ-2017) : доклады XVI Междунар. конф., Минск, 16 нояб. 2017 г. / ОИПИ НАН Беларуси. – Минск, 2017. – С. 299–304.

4. Wav2vec 2.0: A framework for self-supervised learning of speech representations / A. Baevski [et al.] // Advances in Neural Information Processing Systems. – 2020. – Vol. 33. – P. 12449–12460.

5. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks / A. Graves [et al.] // Proc. of the 23rd Intern. Conf. on Machine Learning. – New York, 2006. – P. 369–376.

6. Morris, A. From WER and RIL to MER and WIL: Improved evaluation measures for connected speech recognition / A. Morris, V. Maier, P. Green // Eighth Intern. Conf. on Spoken Language Processing. – Korea, Jeju Island, 2004. – P. 2765–2768.