PRZEGLĄD RUSYCYSTYCZNY 2022, nr 1 (177)

DOI 10.31261/pr.12741

DANIEL BORYSOWSKI
ORCID: https://orcid.org/0000-0001-6594-9047
Uniwersytet Opolski

WYKORZYSTANIE KORPUSÓW ROSYJSKOJĘZYCZNYCH NEWSÓW INTERNETOWYCH NA POTRZEBY SYSTEMÓW AUTOMATYCZNEGO ROZPOZNAWANIA MOWY W OBSZARZE MONITORINGU MEDIÓW

THE USE OF RUSSIAN-LANGUAGE INTERNET NEWS CORPORA FOR THE PURPOSES OF AUTOMATIC SPEECH RECOGNITION SYSTEMS IN THE AREA OF THE MEDIA MONITORING

The author of the article used open Internet-news corpuses (NewsRu and Taiga) to create N-gram language models for the needs of automatic speech recognition systems. The models were comprehensively evaluated (perplexity, WER, proper name recognition, comparison with the base model and Google ASR). The author also rescored N-gram models, using recursive neural networks. The effectiveness of the models was assessed by recognizing speech from the news channel Россия 24 (37 files with a total length of 1.5 hours were tested). The selection of test data is related to the main goal of the article — speech recognition for the needs of the so-called media monitoring.

1. WPROWADZENIE

Istota percepcji mowy jako problem badawczy interesowała naukowców od dziesięcioleci. W wieku XIX ojciec amerykańskiej psychologii, William James, stwierdził: „Gdy słuchamy czyjejś mowy, wiele z tego, co — jak nam się wydaje — słyszymy, pochodzi z naszej pamięci”[1]. Dziś, szczególnie w odniesieniu do automatycznego rozpoznawania mowy, to zdanie nabiera nowego sensu, choć gdy zostało opublikowane, badania w tym obszarze dalekie były od stworzenia faktycznego urządzenia ASR (ang. *automatic speech recognition*). Pierwsze warte odnotowania postępy w instrumentalizacji teorii dotyczącej

[1] W. James, *Talks to Teachers on Psychology: And to Students on Some of Life's Ideals*, Holt, New York, 1889, s. 159. Tłum. — D.B.

odbierania i przetwarzania mowy przez maszyny przypadają na połowę lat 50. XX wieku, kiedy w Bell Labs stworzono analogowy system do rozpoznawania cyfr wypowiadanych przez człowieka do aparatu telefonicznego[2]. W kolejnych dziesięcioleciach metodologia i technologia ASR rozwijane były dzięki postępom w różnych obszarach komputerowego przetwarzania danych[3]. Spośród ośmiu kluczowych obszarów (technologii), wymienionych przez Douglasa O'Shaughnessy'ego, z punktu widzenia niniejszego tekstu szczególne znaczenie ma m o d e l o w a n i e j ę z y k a n a t u r a l n e g o (na poziomie zarówno akustycznym, jak i graficznym[4]).

Obecnie trudno wskazać wyraźne granice między naukowym badaniem ASR, wykorzystaniem ASR w procesach biznesowych (telefoniczna obsługa klientów z wykorzystaniem tzw. botów głosowych, kompleksowa analiza językowa rozmów konsultantów *call centre* z klientami i in.), czy też wykorzystaniem ASR w życiu codziennym (wyszukiwanie głosowe, głosowe zarządzanie urządzeniami typu Smart, automatyczne generowanie napisów i in.)[5]. We wszystkich tych sferach działalności ludzkiej w pewnym stopniu korzysta się z najnowocześniejszych metodologii, zasobów oraz technologii; w ujęciu badawczym wykorzystanie takiego czy innego podejścia, zasobu czy technologii determinuje cel prowadzonych analiz. Jedną z możliwości zastosowania systemów ASR jest tzw. monitoring mediów, czyli rozpoznawanie mowy z programów radiowych i/lub telewizyjnych, nadawanych na żywo i udostępnianych na przykład przez Internet. W niniejszym tekście prezentujemy proces przetwarzania rosyjskojęzycznych danych tekstowych (w różnym stopniu szczegółowości) właśnie na potrzeby modelowania języka oraz rozpoznawania mowy (na przykładzie programu informacyjnego Россия 24[6]).

2. MODELOWANIE JĘZYKA I AUTOMATYCZNE ROZPOZNAWANIE MOWY

Modelowanie języka w ostatnim dziesięcioleciu stało się kluczowym zagadnieniem w ramach przetwarzania języka naturalnego (ang. *na-*

[2] D. O'Shaughnessy, *Invited paper: Automatic speech recognition: History, methods and challenges*, „Pattern Recognition" 2008, nr 41, s. 2966–2967.

[3] Por. tabelę 2. Tamże, s. 2967.

[4] Mamy tutaj na myśli teksty pisane, prymarnie cyfrowe.

[5] Niniejszy artykuł wpisuje się w nurt badań naukowych wspieranych przez sektor komercyjny, powstał bowiem dzięki współpracy autora z firmą VoiceLab.AI Sp. z o. o., która w ramach projektu nr POIR.01.01.01-00–1237/19 otrzymała częściowe dofinansowanie z funduszy Narodowego Centrum Badań i Rozwoju.

[6] Nadawanego na żywo za pośrednictwem strony https://www.vesti.ru/ (10.10.2021).

tural language processing; dalej NLP) i jest wykorzystywane w wielu obszarach związanych, mówiąc najogólniej, z przetwarzaniem tekstów i wyzyskiwaniem z nich jakichś informacji (również jako jeden z subetapów automatycznego rozpoznawania mowy). Obecnie modele języka tworzy się w ramach tzw. uczenia maszynowego z wykorzystaniem sieci neuronowych, choć do różnych zadań z tego obszaru sięgano także po modele N-gramowe. Te ostatnie, zaliczane do modeli statystycznych i deterministycznych, znane były już od lat 50. XX wieku. W praktyce i na większą skalę stosowano je dopiero w kolejnych dekadach. Do dziś bywają wykorzystywane na potrzeby systemów ASR ze względu na łatwość ich budowania i implementowania[7]. Rozwój w sferze uczenia maszynowego doprowadził do wykorzystywania do zadań z obszaru NLP czy ASR modeli RNN, czyli modeli opartych na rekurencyjnych sieciach neuronowych (ang. *recurrent neural networks*). Stanowiły one albo podstawowe narzędzie, albo mechanizm dodatkowy, na przykład poprawiający rezultaty modeli N-gramowych. Ważnym wydarzeniem w sferze NLP było stworzenie algorytmu word2vec[8], również wykorzystującego sieci neuronowe. W 2017 roku z kolei po raz pierwszy zaprezentowano nowy typ modelu — tzw. Transformer — obecnie najczęściej wykorzystywany typ modeli do zadań z obszaru przetwarzania języka naturalnego[9].

W odniesieniu do ASR warto wymienić publikacje dotyczące historii tej technologii[10]. Tematykę ASR szczegółowo omawiają także podręczniki akademickie. Przykładowo w podręczniku Ivana Tam-

---

[7] Więcej informacji na temat modeli N-gramowych por. choćby internetową (stale aktualizowaną i rozbudowywaną od roku 2008) publikację D. Jurafsky, J.H. Martin, *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Third Edition draft, 2021, s. 30–35 https://web.stanford.edu/~jurafsky/slp3/ed3book_sep212021.pdf (24.10.2021).

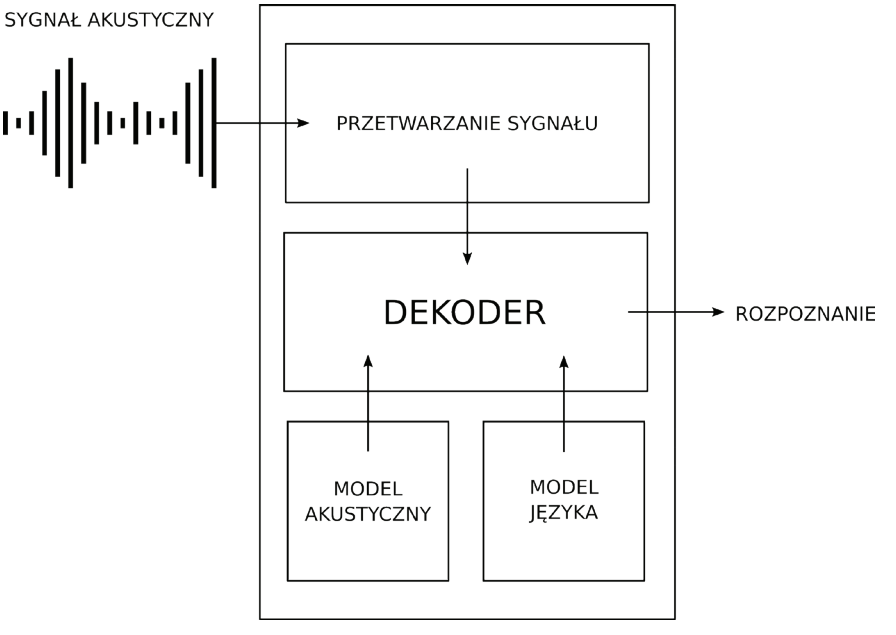
[8] T. Mikolov i in., *Efficient Estimation of Word Representations in Vector Space*, 2013, https://arxiv.org/abs/1301.3781v3 (20.10.2021) oraz T. Mikolov i in., *Distributed Representations of Words and Phrases and their Compositionality*, 2013, https://arxiv.org/abs/1310.4546v1 (20.10.2021).

[9] A. Vaswani i in., *Attention Is All You Need*, 2017, https://arxiv.org/abs/1706.03762v5 (21.10.2021) oraz T. Wolf i in., *Transformers: State-of-the-Art Natural Language Processing*, 2020, https://aclanthology.org/2020.emnlp-demos.6.pdf (21.10.2021).

[10] Zob. między innymi D. O'Shaughnessy, *Invited paper: Automatic speech recognition…*, B. Ziółko, M. Ziółko, *Przetwarzanie mowy*, Wydawnictwa AGH, Kraków,2011 oraz И.Б. Тампель, *Автоматическое распознавание речи — основные этапы за 50 лет*, „Научно-технический вестник информационных технологий, механики и оптики" 2015, vol. 15, nr 6, s. 957–968.

pelya i Alekseya Karpova[11] kwestie automatycznego rozpoznawania mowy rozpatrywane są od etapu powstawania mowy w aparacie artykulacyjnym człowieka, aż po etap komputerowego przetwarzania sygnału dźwiękowego na tekst.

Zwięzłą informację dotyczącą modelowania języka oraz automatycznego rozpoznawania mowy wieńczymy schematem, który przedstawia działanie systemu ASR[12].



Rys 1. Podstawowy schemat systemu ASR

„Na wejściu" do systemu trafia sygnał akustyczny, na przykład w postaci pliku dźwiękowego. Odpowiednio przetworzony sygnał (do sekwencji obserwacji akustycznych) trafia do tzw. dekodera (programu dekodującego). Dzięki modelowi akustycznemu dekoder generuje sekwencję wyrazów odpowiadających otrzymanym reprezentacjom akustycznym. Model języka pomaga w wygenerowaniu naturalnego

[11] И.Б. Тампель, А.А. Карпов, *Автоматическое распознавание речи. Учебное пособие*, Университет ИТМО, Санкт-Петербург 2017.

[12] Schemat zaczerpnięty z R. Justo, O. Saz, A. Miguel, M.I. Torres, E. Lleida, *Improving Language Models in Speech-Based Human-Machine Interaction*, „International Journal of Advanced Robotic Systems" 2013, Vol. 10 (87), s. 1–11. Wykonanie — D.B.

tekstu w języku wyjściowym[13]. Jednym z kluczowych elementów skutecznego rozpoznawania mowy są tzw. słowniki[14]. Słownik to w istocie plik tekstowy, w którym każdy wyraz zajmuje jedną linię i jest podany wraz z jego postacią wyjściową oraz odpowiadającą mu fonemizacją. System ASR rozpoznaje tylko te wyrazy, które znajdą się w słowniku. Dla systemu korzystającego z modelu akustycznego literowego (więcej o tym zob. 4. Opis bazy materiałowej) przykładowe pozycje w słowniku przyjmą następującą postać:

автоматическое [автоматическое] а в т о м а т и ч е с к о е
распознавание [распознавание] р а с п о з н а в а н и е
речи [речи] р е ч и
антона [Антона] а н т о н а

Wyrazy w słowniku — po lewej stronie — mają zazwyczaj taką samą postać, jak wyrazy w tekstowym korpusie, wykorzystanym do modelu (jeśli sprowadziliśmy wszystkie znaki alfabetu do małych liter). Słowa w nawiasach kwadratowych odpowiadają postaci wyjściowej z systemu (tak zostaną zapisane na przykład w pliku wyjściowym, wyświetlone na ekranie itp.). Pojedyncze litery po stronie prawej odpowiadają reprezentacjom akustycznym, które dekoder z rysunku 1 zamienia na wyrazy. System działa, jeśli model akustyczny został wytrenowany również przy użyciu słownika o takiej postaci (więcej o tym zob. 4. Opis bazy materiałowej).

3. MODELOWANIE JĘZYKA ROSYJSKIEGO I ASR DLA JĘZYKA ROSYJSKIEGO — ZARYS PROBLEMÓW

Język rosyjski z kilku powodów stanowi wyzwanie w odniesieniu do jego modelowania (warstwy tekstowej), a także rozpoznawania (warstwy akustycznej). W warstwie tekstowej wyzwania te wiążą się z przynależnością tego języka do grupy języków fleksyjnych[15]. Proble-

[13] Tamże, s. 2.

[14] Są one wkompilowane w model języka.

[15] W przypadku licznych zadań z obszaru NLP przed procesem modelowania języka stosuje się tzw. lematyzację, czyli sprowadzenie wszystkich wyrazów do ich formy podstawowej. Teksty wykorzystane do tworzenia modeli akustycznych oraz modeli języka na potrzeby systemów ASR nie podlegają lematyzacji, aczkolwiek bywa ona stosowana do tzw. modeli dwuskładnikowych, w których równolegle z modelowaniem sekwencji wyrazów modelowane są także sekwencje właściwości gramatycznych tych wyrazów (por. A. Karpov, K. Markov, I. Kipyatkova, D. Va-

matyczna jest też duża dowolność szyku wyrazów w języku rosyjskim, co prowadzi do zamodelowania wielu „nadmiarowych" kontekstów, zmniejszając tym samym skuteczność modelu i zwiększając jego rozmiar. Por. zdania:

[1] Вчера он был на работе.
[2] Он вчера был на работе.
[3] Он был вчера на работе.
[4] Он на работе был вчера.
[5] Он был на работе вчера.
[6] На работе он был вчера.
[7] Вчера на работе был он.
[8] На работе вчера был он.

Dla językoznawcy różnica między tymi zdaniami jest oczywista — choćby ze względu na ich właściwości prozodyczne, wynikające z ich struktury tematyczno-rematycznej. Prawdopodobnie każda osoba władająca językiem rosyjskim bez problemu wychwyci wszelkie niuanse znaczeniowe zawarte w zdaniach [1]–[8]. Dla osób tych będzie oczywiste, że zdania [1]–[3] mówią o tym, że wczoraj pewien „on" był w pracy (a nie na zwolnieniu, urlopie itp.); zdania [4]–[6] informują z kolei o tym, że tenże „on" w pracy był wczoraj (a nie przedwczoraj), zdania [7]–[8] natomiast podkreślają, że w pracy był „on" (a nie kto inny)[16]. Dla modelu języka, którego źródłem jest tekst pisany (cyfrowy), różnicę stanowi jedynie szyk. Oczywiście nie każda z powyższych sekwencji wyrazów zostanie zamodelowana równie dobrze, co zależy od liczby ich wystąpień w przetwarzanym korpusie tekstowym[17].

Inny problem w warstwie tekstowej (ortograficznej) stanowi dywiz. Z jednej strony istnieją w języku rosyjskim wyrazy o stałej lub względnie stałej pisowni z dywizem, na przykład zaimki nieokreślone что-то,

zhenina, A. Ronzhin, *Large vocabulary Russian speech recognition using syntactico-statistical language modeling*, „Speech Communication" 2013, Vol. 56, s. 213–228 oraz I. Kipyatkova, A. Karpov, *Study of Morphological Factors of Factored Language Models for Russian ASR*, w: A. Ronzhin i in. (eds.), *Speech And Computer*, Springer, Switzerland 2014, s. 451–458.

[16] Subtelne różnice semantyczne między poszczególnymi zdaniami występują także w ramach trójek [1]–[3] i [4]–[6], a także w ramach dwójki [7]–[8], ale wymagałoby to precyzyjnego dowodzenia, które dla głównego wywodu nie jest niezbędne.

[17] Odnosimy się tutaj głównie do modeli N-gramowych oraz RNN-owych. Nowsze typy modeli, np. modele Transformerowe (zob. przypis 9), wykorzystują mechanizm tzw. atencji, i w pewnym stopniu potrafią wychwytywać zawiłości tematyczno-rematycznej struktury zdań.

кому-то, где-либо, кое-кто, przysłówki по-русски, по-моему, czy też połączenia typu чёрно-белый, культурно-просветительский, z drugiej zaś — szczególnie w potocznej odmianie ruszczyzny — takie dwukomponentowe twory, jak я-то, работаешь-то, успел-таки, принёс-таки i in. Łączliwość wyrazów z partykułą -то jest w zasadzie nieograniczona (czy też ograniczona jedynie przez uzus). Posiadanie wszystkich możliwych wariantów takich połączeń w słowniku znacznie zwiększy jego rozmiar (por. проблема-то, проблемы-то, проблем-то i in.) oraz może negatywnie wpłynąć na rozpoznanie w niektórych kontekstach.

Ostatnim problemem w warstwie tekstowej jest litera ё, choć stanowi ona także problem w warstwie akustycznej. Jednym z rozwiązań tego problemu jest zastępowanie litery ё literą e i złożenie kwestii dwoistej interpretacji akustycznej tego samego znaku graficznego na karb modelu akustycznego. Na poziomie analizy korpusów również wydaje się to uzasadnione, ponieważ sami Rosjanie raczej nie stosują litery ё w piśmie bądź stosują ją tylko w przypadkach kontekstowo niejednoznacznych (z reguły w tekstach literackich czy publicystycznych). Takie rozwiązanie zastosowaliśmy w naszych eksperymentach, co oznacza, że dla reprezentacji wyrazów всё i все w naszym słowniku znajduje się tylko jedna pozycja o postaci „все [все] в с е", i to model akustyczny „wie", że fonemowi e („fonemowi-literze" w literowym wariancie modelu) mogą odpowiadać w tym wypadku dwie wartości akustyczne.

Wydaje się, że większe wyzwanie stanowi warstwa akustyczna języka rosyjskiego. Ruchomy i swobodny akcent oraz związana z nim redukcja samogłosek w pozycji nieakcentowanej to główny problem systemów ASR służących rozpoznawaniu tego języka. Wyżej przedstawiliśmy fragment słownika, w którym zastosowano tzw. wariant literowy fonemizacji. W praktyce oznacza to znacznie łatwiejsze przygotowanie słownika na potrzeby ASR, lecz jednocześnie bardziej złożony proces tworzenia modelu akustycznego[18]. Wariant fonemowy tworzony jest na podstawie reguł, o ile takie w danym języku istnieją. W wypadku języka rosyjskiego — tak, wartości fonetyczne samogłosek są jednak uzależnione od pozycji akcentu w wyrazie. Rozwiązanie

[18] Modele akustyczne trenuje się na parach plik dźwiękowy + plik tekstowy (z dokładnym zapisem tego, co zostało wypowiedziane w pliku dźwiękowym). W procesie treningu również stosuje się słowniki, które wiążą odpowiednie sekwencje fonemów i odpowiadające im wyrazy z warstwą akustyczną pochodzącą z pliku dźwiękowego.

stanowią programy zwane akcentorami statystycznymi[19], które pozwalają na otrzymanie fonemizacji na przykład o takiej postaci[20]:

автоматическое [автоматическое] A F T A M A To Io TSHo I S K A Jo I
распознавание [распознавание] R A S P A Z N A V Ao No I Jo I
речи [речи] Ro Eo TSHo I
антона [Антона] A N T Oo N A

Ostatnią kwestią wartą omówienia jest twarda bądź miękka wymowa wyrazów z literą „e". Wyraz секс powinien być wymawiany twardo, podobnie wyrazy фонетика, энергия, компьютер i wiele innych. W niektórych przypadkach dopuszcza się oba warianty — miękki i twardy (por. na przykład wyraz бассейн), choć w odniesieniu do działania systemów ASR i tak istotne jest to, w jaki sposób ktoś coś wypowie, a nie to, jaką wymowę sugerują słowniki ortoepiczne (jaka jest przyjęta norma). Wspomniany w przypisie 19. program russian_g2p z tymi wypadkami sobie nie radzi i konsekwentnie oferuje fonemizacje ze zmiękczeniem, co być może statystycznie jest nieistotne i nie wpłynie na błędne rozpoznanie tych wyrazów. W ramach wspólnych działań z firmą VoiceLab.AI (głównie z zespołem zajmującym się modelami akustycznymi) testowaliśmy różne rozwiązania, by ostatecznie pozostać przy wariancie literowym słownika (z fonemizacją zapisaną cyrylicą bez jakichkolwiek dodatkowych oznaczeń)[21].

4. OPIS BAZY MATERIAŁOWEJ

Na początku tej części musimy wyjaśnić, co rozumiemy pod pojęciem „korpus". Nie chodzi tu bowiem o typowy zbiór wyselekcjonowanych tekstów, zazwyczaj anotowanych, wyposażony w mniej lub bardziej zaawansowane narzędzie wyszukujące. W takich przypadkach dostęp do samych składowych korpusu jest ograniczony (teksty korpusu wyświetlane są fragmentarycznie, odpowiednio do określonego w kwe-

[19] Zob. O. Yakovenko, I. Bondarenko, M. Borovikova, D. Vodolazsky, *Algorithms for automatic accentuation and transcription of Russian texts in speech recognition systems*, w: A. Karpov, O. Jokisch, R. Potapova (red.), *Speech And Computer*..., s. 768–777. Por. również https://github.com/nsu-ai/russian_g2p (15.10.2021).

[20] Znak zera przy samogłoskach oznacza akcent, natomiast przy spółgłoskach — obecność bezpośrednio po nich samogłoski zmiękczającej.

[21] Zastosowanie takiego rozwiązania wymaga jednak treningu modelu akustycznego na znacznie większej ilości danych — maksymalnie zróżnicowanych pod względem zarówno jakości, jak i przynależności gatunkowej.

rendzie korpusowej kontekstu). Z punktu widzenia przetwarzania języka naturalnego, a tym samym modelowania języka, potrzebne są po prostu dane tekstowe, na przykład w postaci pojedynczego pliku, którego kolejne linie stanowią swego rodzaju rekordy (każda linia to przykładowo jeden autonomiczny tekst, jedno zdanie, kilka zdań)[22]. To, jakiej postaci dane stosowane są do konkretnego zadania NLP, zależy od jego celu. Cel taki może stanowić klasyfikacja tekstów, generowanie słów kluczowych, tworzenie streszczeń, odtwarzanie interpunkcji, korekta rozpoznań OCR (ang. *optical character recognition*) i wiele innych[23]. W modelowaniu języka pod kątem automatycznego rozpoznawania mowy przez korpus zwykle rozumie się zbiór (kolekcję) tekstów w formie plików tekstowych, takich, które można swobodnie przetworzyć w całości.

Do prezentowanego tutaj eksperymentu wykorzystaliśmy dwie ogólnodostępne kolekcje:

1) Korpus NewsRu (ponad 2,5 mln rosyjskojęzycznych newsów internetowych)[24],

2) Korpus Taiga, a dokładniej jego subkorpus, również dot. newsów internetowych (ponad 470 tys. tekstów)[25].

Pierwsza kolekcja gromadzi newsy internetowe z czterech popularnych serwisów informacyjnych (Вести, Интерфакс, Лента

[22] Inną możliwością jest przechowywanie każdego tekstu/zdania w oddzielnych plikach, jeszcze inną — stosowanie różnych formatów plików służących do przechowywania danych, np. w układzie kolumnowym (TSV, CSV, CoNLL-X/CONLL-U) lub w postaci tzw. obiektów (np. JSON). W niniejszym tekście określeń „korpus tekstów", „kolekcja tekstów", „zbiór tekstów" będziemy używać zamiennie, jednak zawsze w odniesieniu do danych nieanotowanych, bez interfejsu graficznego, czy też bez „wbudowanego" graficznego narzędzia przeszukującego.

[23] Por. zadania w ramach kampanii ewaluacyjnej PolEval — http://poleval.pl/ (15.20.2021). Opis jednego z nowszych i wysoce skutecznych modeli stosowanych obecnie w tego typu zadaniach (tzw. modelu T5), zob. C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P.J. Liu, *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*, „Journal of Machine Learning Research" 2020, vol. 21, s. 1–67.

[24] D. Borysowski, *Web crawling dla celów lingwistycznych. Wybrane aspekty gromadzenia i analizy danych tekstowych na przykładzie rosyjskojęzycznych newsów internetowych*, „Prace Językoznawcze" 2021, vol. XXIII/3, s. 87–104. Zob. także https://osf.io/697md/ (10.10.2021).

[25] T. Shavrina, O. Shapovalova, *To the Methodology of Corpus Construction for Machine Learning: „Taiga" Syntax Tree Corpus and Parser*, w: V.P. Zakharov (red.), *Proceedings of the International Conference „Corpus Linguistics–2017"*, 2017, St. Petersburg, s. 78–84. Zob. także https://tatianashavrina.github.io/taiga_site/ (10.10.2021).

i Фонтанка) z lat 1999–2019. Druga — dane z serwisów Интерфакс, Лента i Фонтанка, a dodatkowo także z serwisu Комсомольская правда, w różnych proporcjach obejmując lata 2010–2017. Konstrukcja każdego korpusu jest nieco inna. Korpus NewsRu dostępny jest jako cztery odrębne pliki w formacie JSON. Fragment takiego pliku prezentujemy poniżej.

{
"metadata": {
"url": "https://lenta.ru/news/2019/12/24/oprovergli/",
"date": „2019–12–24",
"time": "02:40:00",
"category": "Бывший СССР",
"subcategory": "Украина",
"title": "СБУ прокомментировала сообщение о расстреле офицеров в Закарпатье",
"tags": ["Прибалтика", "Белоруссия", "Молдавия", "Закавказье", "Казахстан", "Средняя Азия"]
},
"text": "Пресс-секретарь Службы безопасности Украины (СБУ) Елена Гитлянская опровергла информацию о расстреле оперативной группы в Закарпатье. Об этом она написала в Facebook. «Уважаемые журналисты, не распространяйте фейк о \"расстреле группы СБУ в Закарпатье\". Это полный бред!», — говорится в сообщении представителя СБУ. О пытках и убийстве четырех сотрудников спецслужбы ранее, 23 декабря, сообщило украинское издание «Страна.ua». В публикации говорилось, что офицеров расстреляли в Закарпатье местные бандиты, которые занимались контрабандой. На момент публикации официального подтверждения информации не было."
}

Przy użyciu odpowiedniego programu korpus ten można w dowolny sposób filtrować. Dzięki temu możliwe jest tworzenie rozmaitych podzbiorów danych (z tytułami, tagami, kategoriami, autorami), zliczanie liczby newsów opublikowanych w zadanym okresie (roku, miesiącu, dniu, przedziale godzinowym), zapisywanie wybranych elementów korpusu do plików wyjściowych (na przykład tylko tekstu zasadniczego newsów z wybranych rekordów).

Korpus Taiga dostępny jest w dwóch postaciach. Jako zbiór plików tekstowych z tekstem zasadniczym newsów w formacie .txt — wraz z opisującymi go metaplikami (w formacie .csv), a także jako baza danych sqlite3. Pojedynczy rekord metapliku dotyczącego newsów z serwisu Лента prezentujemy poniżej w formie tabeli.

Tabela 1. Metaplik dotyczący pojedynczego newsa internetowego z zasobu Taiga (Лента)

Pole	Zawartość
segment	Lenta
Textid	20150806moonside
textname	НАСА показало обратную сторону Луны на фоне Земли
textrubric	Космос
Date	6 августа 2015
Time	07:09
Source	https://lenta.ru/news/2015/08/06/moonside/

Powyższy metaplik powiązany jest z plikiem tekstowym 20150806moonside.txt. Por. jego fragment:

> Космическому аппарату НАСА DSCOVR удалось получить редкий снимок обратной стороны Луны на фоне Земли — из точки Лагранжа системы Земля-Солнце (объекты, находящиеся в ней, одинаково притягиваются к обоим космическим телам). Об этом сообщается в пресс-релизе агентства. Исходную серию изображений камера Earth Polychromatic Imaging Camera (EPIC) на борту DSCOVR получила 16 июля 2015 года, с расстояния 1,5 миллиона километров от Земли […].

Wybrane teksty z obu zasobów posłużyły nam do stworzenia dwóch modeli języka, które przetestowaliśmy z wykorzystaniem systemu ASR firmy VoiceLab.AI. Dane testowe stanowią fragmenty programów informacyjnych z kanału Россия 24 o łącznej długości ok. 1,5 godziny (37 plików). Dokonano ręcznej transkrypcji nagrań audio oraz — jako zadanie dodatkowe — ręcznej anotacji nazw własnych z podziałem na siedem kategorii: osoby (190 segmentów), wydarzenia (12 segmentów), produkty (28 segmentów), organizacje (43 segmenty), daty (21 segmentów), adresy (2 segmenty), lokalizacje (245 segmentów)[26]. Podstawowym celem naszego badania było sprawdzenie, czy (i w jakim stopniu) wykorzystanie tych samych danych tekstowych z dwóch różnych źródeł wpływa na jakość modelu języka, a w efekcie — na wyniki automatycznego rozpoznawania mowy. Anotacja danych pozwoliła również na ustalenie skuteczności obu modeli w rozpoznawaniu nazw własnych.

[26] Modele, słowniki oraz dane testowe dostępne są w serwisie OSF na prawach licencji CC BY-NC 4.0 (https://osf.io/a6hm5/). Wśród danych tych czytelnik nie odnajdzie jedynie pliku z modelem akustycznym oraz plików wejściowych dekodera, stanowiących własność firmy VoiceLab.AI.

5. PROCES TWORZENIA MODELI JĘZYKA ORAZ ICH EWALUACJI

Opisane korpusy — dzięki metadanym — mogą być wykorzystywane do różnych zadań z obszaru NLP; nas interesowała jedynie treść zawartych w nich tekstów. Zbiór NewsRu jest kilkukrotnie większy od zbioru Taiga (jego subkorpusu newsów), dlatego do naszych eksperymentów wykorzystaliśmy wszystkie dokumenty tekstowe zbioru Taiga News (z wyłączeniem danych serwisu Комсомольская правда) oraz wszystkie odpowiadające im dokumenty ze zbioru NewsRu (rzecz jasna, z wyłączeniem danych serwisu Вести). Tę odpowiedniość ustaliliśmy dzięki identyczności adresów URL w metadanych obu zbiorów. W sumie dało to około 400 tysięcy pokrywających się „rekordów". Tak przygotowane zasoby zostały podzielone na część treningową, walidacyjną oraz testową w proporcji 0,8/0,1/0,1. Dane treningowe wykorzystaliśmy do stworzenia modeli N-gramowych (a dokładniej — 3-gramowych) oraz modeli RNN[27]. Te drugie miały posłużyć do rescoringu, czyli poprawy wyników uzyskanych z użyciem modeli bazowych[28]. Proces podziału na tzw. zbiory TRAIN, VALID i TEST poprzedziła normalizacja korpusów, która oznaczała odpowiednio:

1) usunięcie znaków interpunkcyjnych (w tym zamianę dywizu na spację),

2) zamianę litery ë na literę e,

3) zamianę wielkich liter na małe,

4) zamianę cyfr i liczb na pomocniczy tag <NUMB>, który z założenia powinien pomóc w lepszym zamodelowaniu kontekstów z liczebnikami, datami i in.,

5) usunięcie znaczników HTML, adresów internetowych oraz adresów poczty elektronicznej,

6) usunięcie nie-cyrylicy.

Po zastosowaniu normalizacji przytaczany już fragment tekstu o działaniach organizacji NASA przyjął następującą postać:

> космическому аппарату наса удалось получить редкий снимок обратной стороны луны на фоне земли из точки лагранжа системы земля солнце объ-

[27] Modele N-gramowe zbudowaliśmy narzędziem IRSTLM (zob. https://hlt-mt.fbk.eu/technologies/irstlm; 26.10.2021), modele RNN natomiast — narzędziem opartym na kodzie TensorFlow (zob. https://www.tensorflow.org/; 26.10.2021).

[28] Por. wyniki badań prowadzonych w tym kierunku: I. Kipyatkova, A. Karpov, *Recurrent Neural Network-based Language Modeling for an Automatic Russian Speech Recognition System*, w: S. Balandin i in. (red.), *Proceedings of AINL-ISMW FRUCT Conference*, St. Petersburgs 2015, s. 33–38.

екты находящиеся в ней одинаково притягиваются к обоим космическим телам об этом сообщается в пресс релизе агентства исходную серию изображений камера на борту получила <NUMB> июля <NUMB> <NUMB> года с расстояния <NUMB> <NUMB> миллиона километров от земли

Modele N-gramowe będziemy dalej nazywać — zgodnie z formatem, w jakim są zapisane — plikami ARPA[29], dodając prefiks danego zasobu, na przykład NewsRu ARPA oraz Taiga ARPA. Modele RNN będziemy odpowiednio identyfikować nazwami NewsRu RNN, Taiga RNN. Określenie „model języka" zwyczajowo stosuje się zarówno w odniesieniu do formatu ARPA, jak i w odniesieniu do modeli RNN, a także finalnego pliku wejściowego do dekodera, dlatego — w celu uniknięcia terminologicznych nieporozumień — wprowadzamy powyższe rozróżnienie. Dodatkowo gotowy system automatycznego rozpoznawania mowy będziemy oznaczać skrótem „ASR". Podstawą w przeprowadzonym eksperymencie będą dla nas zatem wszystkie wymienione modele, a także systemy NewsRu ASR, Taiga ASR, punktem odniesienia natomiast — Base ASR[30] oraz Google ASR. Ewaluacja plików ARPA, modeli RNN oraz kompletnych systemów ASR odbywała się na dwóch poziomach:

1) Perplexity (ARPA i RNN) na danych VALID i TEST oraz na plikach tekstowych właściwego zbioru testowego (dalej TEST SET)[31],

2) WER (ASR) na danych TEST SET[32].

Wyniki analizy perplexity prezentujemy w tabeli 2. Tabela zawiera również informację o tzw. OOV (ang. *out-of-vocabulary*), czyli liczbie wyrazów z danych testowych i walidacyjnych, których nie ma w modelu (wyrażonej procentowo). Pogrubieniem oznaczamy naj-

[29] M. Federico, N. Bertoldi, M. Cettolo, *IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models*, w: *Proceedings of Interspeech*, Brisbane 2008, s. 1618–1621.

[30] Base ARPA oraz Base RNN zostały sworzone z danych korpusu OpenSubtitles 2018 (https://opus.nlpl.eu/; 29.10.2021), wybranych danych z korpusu Taiga oraz danych komercyjnych kupionych przez firmę VoiceLab.AI.

[31] Zob. D. Jurafsky, J.H. Martin, *Speech and Language Processing*..., s. 36–37. Mówiąc najogólniej, perplexity jest miarą określającą skuteczność przewidywania przez model języka danych testowych, których wcześniej „nie widział", czyli nie został stworzony na ich podstawie. Im niższa wartość perplexity, tym lepszy model.

[32] Od ang. *word error rate*. Jest to miara określająca, jaki procent rozpoznania stanowią błędy. Błędy te z kolei możemy podzielić na trzy kategorie: usunięcia (ang. *deletions*), wstawienia (ang. *insertions*) oraz podstawienia (ang. *substitutions*). Suma wszystkich usunięć, wstawień i podstawień obliczona poprzez porównanie pliku z dokładną transkrypcją oraz pliku z wyjściowym rozpoznaniem z dekodera stanowi wskaźnik WER.

lepszy wynik, przy czym bierzemy pod uwagę tylko modele NewsRu i Taiga.

Tab. 2. Perplexity dla modeli NewsRu, Taiga oraz Base (ARPA i RNN)

Dane	Modele	NewsRu		Taiga		Base	
		Typ modelu		Typ modelu		Typ modelu	
		ARPA	RNN	ARPA	RNN	ARPA	RNN
NewsRu	VALID	PP=411.71 OOV=2.66%	PP=204.76	PP=468.97 OOV=2.72%	PP=237.60	PP=420.42 OOV=2.40%	PP=330.64
	TEST	PP=522.36 OOV=3.03%	PP=234.75	PP=551.76 OOV=3.05%	PP=255.16	PP=425.58 OOV=2.30%	PP=356.69
Taiga	VALID	PP=159.81 OOV=2.39%	PP=199.95	PP=471.89 OOV=2.62%	PP=180.25	PP=505.71 OOV=2.66%	PP=302.59
	TEST	PP=579.77 OOV=3.55%	PP=238.67	PP=599.01 OOV=3.57%	PP=248.13	PP=520.51 OOV=2.84%	PP=354.20
TEST SET (**Россия 24**)		PP=2416.37 OOV=4.17%	PP=692.62	PP=2457.19 OOV=4.21%	PP=719.52	PP=1507.52 OOV=1.65%	PP=796.61

Bardzo niska wartość perplexity dotycząca modelu NewsRu ARPA oraz danych Taiga VALID najprawdopodobniej spowodowana jest wysokim stopniem pokrywania się tych danych z danymi TRAIN modelu NewsRu (podczas procesu przygotowywania korpusów do treningu zasoby dzielone są w sposób losowy). Dla pozostałych zasobów perplexity utrzymuje się na względnie stałym poziomie, co oznacza, że tę jedną wartość należałoby potraktować jako nieistotną.

Kolejnym etapem tworzenia systemu ASR jest opracowanie słownika. Listy słów do słowników są generowane automatycznie podczas procesu obróbki danych korpusowych. Przetwarza się je do list rangowych, aby możliwe było wyselekcjonowanie najczęstszych wyrazów oraz odrzucenie tych najrzadszych bądź po prostu niechcianych. Przygotowując dane do treningu modelu RNN (te same dane zostały potem wykorzystane do stworzenia plików ARPA), wyznaczyliśmy granicę powtórzeń wyrazów na dość wysokim poziomie 48, co dało w wypadku obu zasobów (NewsRu i Taiga) około 130 tysięcy wyrazów. Listy te zostały wykorzystane jako podstawa słowników. Dodatkowo z danych treningowych stworzyliśmy słowniki zawierające po ok. 300 tysięcy wyrazów oraz wszystkie wyrazy z danego zasobu (NewsRu — 642 tysiące, Taiga — 677 tysięcy), aby sprawdzić wpływ

rozmiaru słownika na WER. To oznacza, że do testów zbudowaliśmy sześć modeli (po trzy na podstawie każdego z zasobów z trzema różnymi słownikami).

Jednym z najważniejszych elementów systemu ASR jest model akustyczny. Do testów wykorzystaliśmy model literowy, wytrenowany na 2539 godzinach nagrań[33].

Etap testowania modeli na drugim poziomie — dotyczącym WER — odbywał się w kilku subetapach oraz dotyczył wyłącznie danych TEST SET, czyli 1,5 h nagrań z kanału informacyjnego Россия 24. Kolejność działań była następująca:

1) poszukiwanie najlepszej konfiguracji modelu języka / parametrów modelu akustycznego i dekodera / rozmiaru słownika (modele NewsRu ASR i Taiga ASR),

2) wybór najlepszej konfiguracji,

3) testy modelu Base ASR z użyciem kilku najlepszych wariantów konfiguracji,

4) testy modelu Google ASR (bez możliwości ingerencji w konfigurację, ale z różnymi wariantami normalizacji rozpoznań, zob. przypis 33),

5) rescoring modeli N-gramowych (ARPA) z użyciem modeli RNN,

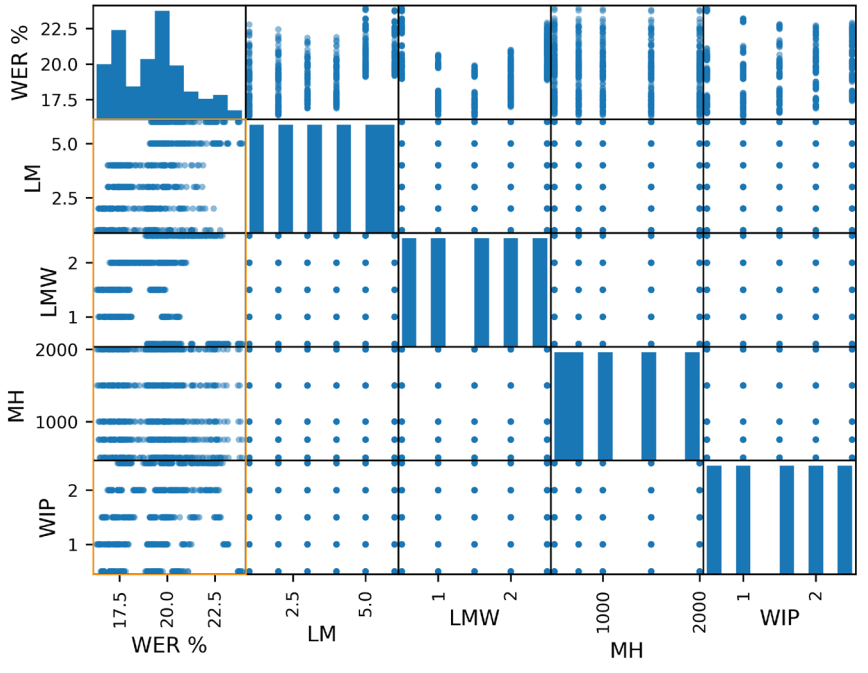
6) testy modeli z rescoringiem dla wybranych konfiguracji.

Dodatkowym subetapem był wspomniany już test poprawności rozpoznawania nazw własnych. Wszystkie powyższe etapy ewaluacyjne zostały przeprowadzone w wariancie lowercase (czyli bez uwzględniania wielkości liter), choć nasze modele są przygotowane do wyświetlania rozpoznań z zachowaniem wielkich liter w jednowyrazowych nazwach własnych.

W pierwszym subetapie sprawdziliśmy 750 wariantów konfiguracji (po 125 dla każdego z sześciu modeli). Konfiguracje obejmowały

[33] Konkretniej są to dane Golos — 1221 godzin (https://github.com/sberdevices/golos; 28.10.2021), 5-procentowa próbka danych Open-STT — 657 godzin (https://github.com/snakers4/open_stt; 28.10.2021), kanał informacyjny Первый — 11 godzin (własne transkrypcje), kanał informacyjny Россия 24 — 9 godzin (dane nieużyte w danych TEST SET; własne transkrypcje), Common Voice — 104 godziny (https://commonvoice.mozilla.org/pl; 28.10.2021), Kaggle — 134 godziny (https://www.kaggle.com/tapakah68/audio-dataset; 28.10.2021), LibriSpeech — 93 godziny (https://www.openslr.org/96/; 28.10.2021), Radio Svoboda — 24 godziny (własne transkrypcje), Multilingual TEDx — 40 godzin (https://www.openslr.org/100; 28.10.2021), Caito — 46 godzin (https://www.caito.de/; 28.10.2021), VoxForge — 18 godzin (http://www.voxforge.org/ru; 28.10.2021) oraz inne dane firmy VoiceLab.AI (pozostałe 182 godziny).

takie parametry, jak LMW (ang. *language model weight*; waga modelu języka względem modelu akustycznego), MH (ang. *max hypotheses*; liczba hipotez z modelu akustycznego), WIP (ang. *word insertion penalty*; kara za wstawianie słów, ograniczająca rozbijanie wyrazów na mniejsze komponenty, jeśli te stanowią samodzielne wyrazy, na przykład разводить/раз водить). Poniżej prezentujemy wyniki ewaluacji[34].



Rys. 2. Zależność WER od rozmiaru słownika oraz parametrów LMW, MH, WIP

Rysunek 2 jasno pokazuje, że najmniejszy wpływ na jakość rozpoznań miał parametr MH, choć nieznacznie lepsze wyniki dawały jego wartości na poziomie 1500–2000. Najlepiej radziły sobie modele 1 i 2 (Taiga ASR ze słownikiem 677 tys. wyrazów oraz NewsRu ASR ze słownikiem 642 tys. wyrazów), choć trudno tu mówić o wyraźnej przewadze (w pierwszych 75 konfiguracjach — WER poniżej progu 17% — pojawiają się także modele 3 i 4 ze średnimi słownikami, a różnica między najlepszym a najgorszym wynikiem w tym przedziale to

[34] Są one także dostępne w formacie .xlsx na stronie https://osf.io/a6hm5/ (29.10.2021).

0,68363 punktu procentowego). Najlepszy wynik WER modelu 5 to 19,08965% (315 pozycja spośród wszystkich konfiguracji; model NewsRu ASR), natomiast modelu 6 — 19,16753% (325 pozycja; model Taiga ASR). Zależność między rozmiarem słownika a wynikiem WER jest bardzo wyraźna. Z rysunku 2 wynikają także najlepsze warianty parametrów LMW i WIP (odpowiednio 1–1,5 i 1). Na 14 najlepszych wyników parametr WIP=1 miały wszystkie, parametr LMW=1,5 — 9 na 14. 5 pozostałych wyników dotyczyło LMW=1. Parametr MH, jak wspomnieliśmy, miał tutaj najmniejsze znacznie, gdyż z wyjątkiem jego wartości równej 500 pozostałe wystąpiły po trzy razy na 14. Najniższy WER odnotował model Taiga ASR — 16.31187% (także dwa kolejne wyniki: 16.32918% i 16.39841%). Model NewsRu ASR zajął czwarte miejsce z wynikiem 16.40706%[35]. Warto zaznaczyć, że różnice na tym poziomie są raczej niezauważalne dla ludzkiego oka i nie sposób ich dostrzec podczas wzrokowej analizy tekstów rozpoznań. W tabeli 3 zamieszczamy listę dziesięciu najczęstszych wyrazów z kategorii usunięć, wstawień i podstawień dla obu porównywanych tutaj zasobów (NewsRu i Taiga) oraz modelu referencyjnego Base ASR. Wzięliśmy pod uwagę tylko najlepsze wyniki.

Tab. 3. Lista najczęstszych wyrazów z kategorii usunięć, wstawień i podstawień

Usunięcia			**Wstawienia**			**Podstawienia**		
NewsRu 1.973%	**Taiga 1.91243%**	**Base 2.11146%**	**NewsRu 3.4441%**	**Taiga 3.51333%**	**Base 3.01142%**	**NewsRu 10.98996%**	**Taiga 10.88612%**	**Base 10.15922%**
и (42)	и (43)	и (41)	в (12)	в (14)	в (13)	ну / но (9)	ну / но (8)	ни / не (5)
в (33)	в (34)	в (38)	и (11)	и (11)	и (12)	ни / не (6)	ни / не (7)	ну / но (5)
а (24)	а (25)	а (16)	на (6)	на (6)	на (7)	из / и (5)	нету / нет (4)	и / таки (4)
не (12)	с (11)	не (12)	это (6)	это (5)	а (5)	нету / нет (4)	в / и (3)	владимир / владимира (3)
с (12)	вот (10)	с (11)	но (4)	а (4)	но (5)	владимир / владимира (3)	владимир / владимир (3)	гуайдо / до (3)
вот (10)	не (9)	вот (7)	а (3)	за (4)	это (5)	которые / который (3)	из / и (3)	с / из (3)

[35] Wyniki te porównaliśmy z wynikami modeli referencyjnych: modelu Base ASR oraz Google ASR. Pierwszy — z parametrami MH=1500, LMW=1,5, WIP=1 — uzyskał wynik WER 15.2821%, drugi — 27.890% (teksty oryginalnych rozpoznań), 24.957% (teksty po zamianie dywizu na spację oraz litery ë na literę e), 22.975% (teksty po normalizacji liczb zapisanych cyframi do zapisu słownego; ASR Google generuje rozpoznania różniące się od przyjętej przez nas normy, dlatego wymagane było ich ujednolicenie przed obliczeniem WER).

у (9)	о (9)	на (6)	за (3)	о (3)	все (3)	не / они (3)	не / они (3)	что / чтобы (3)
на (8)	у (9)	о (6)	о (3)	с (3)	за (3)	с / из (3)	с / из (3)	шарыга / га (3)
о (8)	на (7)	у (6)	с (3)	больше (2)	больше (2)	эта / это (3)	эта / это (3)	эта / это (3)
к (7)	но (7)	как (5)	больше (2)	все (2)	был (2)	акции / акций (2)	академия / академии (2)	академия / академии (2)

W kategoriach usunięć oraz wstawień najczęściej pojawiają się krótkie wyrazy. Operowanie parametrem WIP prowadzi często do przesunięć pomiędzy tymi kategoriami błędów (im wyższy WIP, tym więcej usunięć, im niższy — tym więcej wstawień), jednak nie eliminuje problemu. Nawet odnalezienie optymalnej konfiguracji stanowi raczej kompromis między liczbą usunięć a wstawień. Jeśli chodzi o podstawienia, w tabeli 3 odnajdujemy co najmniej trzy ich przyczyny:

1) podobieństwo na poziomie fonetycznym związane głównie z redukcją samogłosek (por. не / ни, ни / не, нету / нет, это / эта, ну / но);

2) problemy z nazwami własnymi — na przykład z ich brakiem w słowniku, przez co model podstawia w ich miejsce wyrazy o zbliżonej reprezentacji akustycznej (por. гуайдо / до, шарыга / га poprzedzone innymi wyrazami zbliżonymi w warstwie dźwiękowej do гуай / шары);

3) ewentualne błędy w transkrypcjach, stanowiących podstawę porównania z rozpoznaniami (por. владимир / владимира, акции / акций, академия / академии).

Właściwie wszystkie opisane tutaj kategorie mogą dotyczyć po prostu nieprawidłowości w ręcznych transkrypcjach, które zawsze obarczone są błędem ludzkim. Problem z punktu 2) jest jednym z podstawowych wyzwań współczesnych systemów automatycznego rozpoznania mowy. Sprawdzenie jakości naszych modeli w tym obszarze było zresztą naszym celem dodatkowym. Nazwy własne z siedmiu zdefiniowanych przez nas kategorii obejmowały łącznie 6,72% tekstów transkrypcji. Oba modele poradziły sobie z tym zadaniem na zbliżonym poziomie skuteczności — model NewsRu w 74,68%, model Taiga — w 75,1% (przypomnijmy, że analiza dotyczyła wariantu lowercase). Wyżej wzmiankowaliśmy, że analizowane modele są przystosowane do wyświetlania jednowyrazowych nazw własnych. Aby sprawdzić ich działanie, wykonaliśmy jeszcze jeden dodatkowy test, tym razem w wariancie uppercase. Poniżej prezentujemy tekst roz-

poznany przez oba porównywane tu systemy z użyciem największych słowników. Błędnie rozpoznane wyrazy oznaczamy odpowiednimi numerami w nawiasach kwadratowych — [1] NewsRu ASR: ogółem uppercase WER 17.70509% (w poniższym przykładzie — 11.88118%), [2] Taiga ASR: ogółem uppercase WER 17.67047% (w poniższym przykładzie — 11.88118%). Dla czytelności znakiem „/" oznaczamy granice zdań.

> сенат США не смог одобрить законопроект по санкциям против Сирии и ее союзников Ирана и России / документ даже не дошел до голосования не набрав двух т р е т ь е й [1 i 2 zamiast третей] голосов / рассмотрение перенесли / законопроект предусматривает санкции за военную или иную помощь правительству Сирии и даже за сотрудничество в производстве углеводородов / прописанные знакомые обвинения в адрес России к о т о р о е [1 zamiast которая] якобы мешала доставки гумпомощи что н а ш и [1 i 2 (наша) zamiast наше] Минобороны не раз о п р о в е р г а л а [1 i 2 zamiast опроверагло] / впрочем дело не в содержании бумаги а в американском ш а н д а у [1 i 2 zamiast шатдауне] н е к о т о р ы е [1 i 2 zamiast который] тянется с конца декабря / демократы в сенате не желают рассматривать никакие инициативы кроме тех что п о л о ж и т [1 i 2 zamiast положат] конец ш а т у н у [1 i 2 (шандау ну) zamiast шатдауну] и п о з в о л и т [1 i 2 zamiast позволят] правительству США получить финансирование и возобновить

Przedstawienie kompletnych wyników w tym miejscu nie jest możliwe, jednak już jeden przytoczony tekst obrazuje zarówno mocne, jak i słabe strony stworzonych przez nas modeli (a także ogólne problemy związane z rozpoznawaniem mowy w języku rosyjskim). Podstawowym problemem, wymienianym już wielokrotnie, jest redukcja rosyjskich samogłosek. Wartości fonetyczne zakończeń wyrazów — na przykład нарушение/нарушения, которое/которая i wielu innych — są zbliżone, co, na przykład w połączeniu z szybkim tempem mowy, znacznie utrudnia prawidłowe rozpoznanie. Być może słowniki stworzone na podstawie reguł g2p nie będą obciążone tą bliskością, jednak nie dysponujemy wariantem fonemowego modelu akustycznego, aby móc sprawdzić te różnice. Dysponujemy natomiast wynikami poprzedniego modelu akustycznego firmy VoiceLab.AI (158 godzin), który na danych Россия 24 osiągał WER na poziomie ok. 35%. Pokazuje to, jak kosztowne jest poprawianie wyników ASR (10 razy więcej godzin dało poprawę o ok. 18 punktów procentowych).

Innym problemem jest rozpoznawanie nazw własnych (ok. 75% skuteczności w wariancie lowrcase), choć w omawianym fragmencie tekstu wszystkie nazwy zostały rozpoznane prawidłowo. Nie został rozpoznany wyraz stosowany dla określenia wstrzymania prac ame-

rykańskiego rządu, czyli шатдаун (ang. *shutdown*). Jak widać, modele „próbują" jakoś tę kwestię rozwiązać, wstawiając najbardziej zbliżony wariant brzmieniowy (na podstawie wyrazów zawartych w słowniku). W słownikach znalazły się takie wyrazy jak шандау (od Бад-Шандау, miasto-uzdrowisko w Niemczech) oraz шатуна (dopełniacz wyrazu шатун), które „upodobniły" się do wyrazów шатдауне ([шандау + не]которые) i шатдауну (шатуну i [шандау + ну]). Pierwszy przypadek doprowadził do dwóch błędów — nierozpoznania wyrazu шатдаун oraz niepoprawnego rozpoznania wyrazu который.

Dwa ostatnie subetapy naszych eksperymentów to rescoring modeli N-gramowych oraz testy modeli po rescoringu. Po przeprowadzeniu kilku treningów oraz doborze parametrów uzyskaliśmy nieznaczną poprawę, dodatkowo tylko przy użyciu do rescoringu modelu Base RNN (co obrazuje tabela 4). NewsRu RNN oraz Taiga RNN dawały minimalną poprawę lub wręcz nieco pogarszały poziom WER[36].

Tab. 4. Wyniki rescoringu modeli N-gramowych

Model	WER	Usunięcia	Wstawienia	Podstawienia	Waga rescoringu
[1]	16.06957%	1.81724%	3.69505%	10.55729%	0.1
[2]	16.21668%	1.79128%	3.78159%	10.64382%	
[3]	16.36379%	1.84320%	3.72966%	10.79093%	
[4]	16.44168%	1.79993%	3.76428%	10.87747%	
[B]	14.9187%	1.84320%	3.22776%	9.84770%	

Model [1] to NewsRu ARPA po rescoringu modelem Base RNN, model [2] — to Taiga ARPA po rescoringu tym samym modelem. Modele [3] i [4] to odpowiednio modele NewsRu ARPA i Taiga ARPA po rescoringu odpowiadającymi im modelami RNN. Powyższe wyniki mogą świadczyć o tym, że dla większej poprawy należałoby wykorzystać więcej danych, ale być może nie danych o charakterze ogólnym (por. podstawę modelu Base RNN w przypisie 30), lecz wyłącznie danych domenowych (na przykład publicystycznych). Sam model Base RNN nie wyczerpuje możliwości treningowych, o czym świadczy wysokie perplexity uzyskane na danych testowych tego modelu (ponad

[36] Wyjściowy WER modelu NewsRu ARPA wyniósł 16.40706%, natomiast Taiga ARPA — 16.31187%. Dane treningowe oraz pliki konfiguracyjne udostępniamy w serwisie OSF (https://osf.io/a6hm5/; 30.10.2021).

780), chociaż dla wyniku WER 15,2821%, uzyskanego na modelu Base ARPA, po rescoringu modelem RNN wytrenowanym na tych samych danych uzyskaliśmy spadek WER do 14,91866% (zob. model [B] w tabeli 4). Rescoring modeli N-gramowych NewsRu ARPA i Taiga ARPA z wykorzystaniem tych samych danych do modelu RNN nie przyniósł zadowalającej poprawy[37].

6. PODSUMOWANIE

Skuteczność systemów ASR jest coraz większa, a na ogólny jej wzrost składa się wiele czynników. Na poziomie akustycznym bada się możliwości ograniczenia ilości danych dźwiękowych, których dokładne transkrybowanie jest bardzo kosztowne[38]. Zarówno na poziomie akustycznym, jak i na poziomie graficznym (tekstowym) stosuje się tzw. augumentację danych, czyli wykorzystywanie nieco zmienionych kopii posiadanych danych (lub danych generowanych automatycznie na podstawie posiadanych)[39]. Zapewne kolejnym krokiem w uzyskaniu poprawy w naszym przypadku będzie wykorzystanie któregoś z powyższych podejść. Pierwszym zadaniem w ramach monitoringu mediów może być lepszy dobór danych na poziomie korpusowym. Oba przedstawione tutaj zbiory tekstów (NewsRu i Taiga) dają takie możliwości dzięki bogatemu metaopisowi (tytuły, kategorie, tagi itp.). Problemem może być rozbieżność funkcjonalna między tekstem pisanym a mówionym, dlatego wykorzystanie również techniki wav2vec mogłoby przynieść poprawę rezultatów, stanowi to jednak kolejny problem badawczy, wymagający omówienia w innym miejscu.

---

[37] W jednym ze znanych nam eksperymentów dotyczących tej samej kwestii badacze uzyskali poprawę o 3,67 punktu procentowego, co oznaczało obniżenie WER z poziomu 26,54% do poziomu 22,87%. Por. I. Kipyatkova, A. Karpov, *Recurrent Neural Network-based Language Modeling*..., s. 37.

[38] Por. zastosowanie algorytmu wav2vec, który pozwala na wykorzystanie od zaledwie kilku do kilkunastu godzin nagrań z transkrypcjami i trenowanie modelu akustycznego wyłącznie na danych dźwiękowych, S. Schneider, A. Baevski, R. Collobert, M. Auli, *wav2vec: Unsupervised Pre-training for Speech Recognition*, 2019 (https://arxiv.org/abs/1904.05862v4; 31.10.2021).

[39] A. Laptev, R. Korostik, A. Svischev, A. Andrusenko, I. Medennikov, S. Rybin, *You Do Not Need More Data: Improving End-To-End Speech Recognition by Text-To-Speech Data Augmentation*, 2020 (https://arxiv.org/abs/2005.07157v2; 31.10.2021).

REFERENCES

Borysowski, Daniel. "Web crawling dla celów lingwistycznych. Wybrane aspekty gromadzenia i analizy danych tekstowych na przykładzie rosyjskojęzycznych newsów internetowych." *Prace Językoznawcze*, 2021, Vol. XXIII/3: 87–104.

Federico, Marcello, and Bertoldi, Nicola, and Cettolo, Mauro. *IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models. Proceedings of Interspeech*, Brisbane, 2008: 1618–1621.

James, William. *Talks to Teachers on Psychology: And to Students on Some of Life's Ideals*. New York: Holt, 1889.

Jurafsky, Dan, and Martin, James H. *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Third Edition draft, 2021 <https://web.stanford.edu/~jurafsky/slp3/ed3book_sep212021.pdf>.

Justo, Raquel, and Saz, Oscar, and Miguel, Antonio, and Torres, M.I., and Lleida, Eduardo. "Improving Language Models in Speech-Based Human-Machine Interaction." *International Journal of Advanced Robotic Systems*, 2013, vol. 10 (87): 1–11 <https://www.researchgate.net/publication/258225996_Improving_Language_Models_in_Speech-Based_Human-Machine_Interaction>.

Karpov, Alexey, and Markov, Konstantin, and Kipyatkova, Irina, and Vazhenina, Daria, and Ronzhin, Andrey. "Large vocabulary Russian speech recognition using syntactico-statistical language modeling." *Speech Communication*, 2013, Vol. 56: 213–228.

Kipyatkova, Irina and Karpov, Alexey. "Study of Morphological Factors of Factored Language Models for Russian ASR." *Speech And Computer*. Eds. Ronzhin, Andrey et al. Switzerland: Springer, 2014, 451–458.

Kipyatkova, Irina, and Karpov, Alexey. "Recurrent Neural Network-based Language Modeling for an Automatic Russian Speech Recognition System." *Proceedings of AINL-ISMW FRUCT Conference*. Eds. Balandin, Sergey et al. St. Petersburg, 2015, 33–38.

Laptev, Aleksandr, and Korostik, Roman, and Svischev, Aleksey, and Andrusenko, Andrei, and Medennikov, Ivan, and Rybin, Sergey. *You Do Not Need More Data: Improving End-To-End Speech Recognition by Text-To-Speech Data Augmentation*, 2020 <https://arxiv.org/abs/2005.07157v2>.

Mikolov, Tomas et al. *Distributed Representations of Words and Phrases and their Compositionality*, 2013 <https://arxiv.org/abs/1310.4546v1>.

Mikolov, Tomas et al. *Efficient Estimation of Word Representations in Vector Space*, 2013 <https://arxiv.org/abs/1301.3781v3>.

O'Shaughnessy, Douglas. "Invited paper: Automatic speech recognition: History, methods and challenges." *Pattern Recognition*, 2008, 41: 2966–2967 <https://www.sciencedirect.com/science/article/abs/pii/S0031320308001799>.

Raffel, Collin, and Shazeer, Noam, and Roberts, Adam, and Lee, Katherine, and Narang, Sharan, and Matena, Michael, and Zhou, Yanqi, and Li, Wei, and Liu, Peter J. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." *Journal of Machine Learning Research*, 2020, Vol. 21: 1–67.

Schneider, Steffen, and Baevski, Alexei, and Collobert, Ronan, and Auli, Michael. *wav2vec: Unsupervised Pre-training for Speech Recognition*, 2019 <https://arxiv.org/abs/1904.05862v4>.

Shavrina, Tatiana, and Shapovalova, Olga. "To the Methodology of Corpus Construction for Machine Learning: 'Taiga' Syntax Tree Corpus and Parser." *Proceedings of the International Conference „Corpus Linguistics–2017"*. Zakharov, Viktor Pavlovich. Khokhlova, Mariya Vladimirovna (eds.). St. Petersburg, 2017, 78–84.

Tampel', Ivan Borisovich, and Karpov, Aleksey Anatol'yevich. *Avtomaticheskoye raspoznavaniye rechi. Uchebnoye posobiye*. Sankt-Peterburg: Universitet ITMO, 2017 [Тампель, Иван Борисович, and Карпов, Алексей Анатольевич. *Автоматическое распознавание речи. Учебное пособие*. Санкт-Петербург: Университет ИТМО, 2017].

Tampel', Ivan Borisovich. "Avtomaticheskoye raspoznavaniye rechi — osnovnyye etapy 50 za let." *Nauchno-tekhnicheskiy vestnik informatsionnykh tekhnologiy, mekhaniki i optiki*, 2015, Vol. 15, no. 6: 957–968 [Тампель, Иван Борисович. "Автоматическое распознавание речи — основные этапы за 50 лет." *Научно-технический вестник информационных технологий, механики и оптики*, 2015, Vol. 15, no. 6: 957–968].

Vaswani, Ashish (et al.). *Attention Is All You Need*, 2017 <https://arxiv.org/abs/1706.03762v5>.

Wolf, Thomas (et al.). *Transformers: State-of-the-Art Natural Language Processing*, 2020 <https://aclanthology.org/2020.emnlp-demos.6.pdf>.

Yakovenko, Olga, and Bondarenko, Ivan, and Borovikova, Mariya, and Vodolazsky, Daniil. "Algorithms for automatic accentuation and transcription of russian texts in speech recognition systems." *Speech And Computer*. Eds. Karpov, Alexey, and Jokisch, Oliver, and Potapova, Rodmonga. Switzerland: Springer, 2018, 768–777.

Ziółko, Bartosz, and Ziółko, Mariusz. *Przetwarzanie mowy*. Kraków: Wydawnictwa AGH, 2011.