УДК 004.91
DOI: 10.31673/2412-9070.2020.066163

К. О. ГОРДІЄНКО, студентка;
А. Б. КОБА, ст. викладач;
Т. П. ДОВЖЕНКО, доцент,
Державний університет телекомунікацій, Київ

ДОСЛІДЖЕННЯ СИСТЕМ РОЗПІЗНАВАННЯ ТЕКСТУ ТА ВИЛУЧЕННЯ ДАНИХ З УКРАЇНОМОВНИХ ДОКУМЕНТІВ

Розглянуто наявне програмне забезпечення, основним завданням якого є вилучення інформації з оцифрованих документів. З усього програмного забезпечення відбиралося таке, що ґрунтується на технологіях нейронних мереж та глибокого навчання. Вилучення інформації з документів може відбуватися із застосуванням ручної праці операторів персональних комп'ютерів, що потребує багато часу і не виключає вплив людського фактора, а також оцифруванням документів із подальшим обробленням у програмному забезпеченні, яке ґрунтується на принципі підпорядкування документів шаблонам та правилам, що може впливати на швидкість оброблення даних і необхідність вносити зміни до налаштувань через зміну типу документа.

У статті поставлено завдання дослідити наявне програмне забезпечення для вилучення даних із цифрових документів, засноване на технології нейронних мереж, та їх застосовність до україномовних документів. Для цього було створено простий набір рахунків-фактур, які завантажувались у систему.

Розроблення системи для вилучення інформації з оцифрованих україномовних документів за допомогою нейронних мереж пришвидшить оброблення даних, надасть можливість для їх опрацювання залежно від сфери діяльності користувача цього програмного забезпечення.

Визначено, що сьогодні немає систем, котрі можуть самостійно визначати, які дані необхідні для вилучення з україномовних документів. Наявні системи потребують створення програмного забезпечення, що відіграватимуть роль обкладинки для функціонала систем, які передають свою інформацію через REST API. Обґрунтовано, що найкращою системою є Google Form Parser, проте вона потребує постійного підімкнення до мережі Інтернет, що може стати серйозною перепоною для використання такого продукту в певних сферах діяльності.

Ключові слова: оптичне розпізнання символів; нейронні мережі; глибоке навчання; машинне навчання; вилучення даних.

ВСТУП

У всіх сферах діяльності рано чи пізно настає момент, коли потрібно обробити документи, що існують у фізичному вигляді. Це може бути архівна документація, виготовлена ще до масового поширення персональних комп'ютерів, або навіть документи, створені у сьогоденні, які з якихось причин було відправлено через поштові служби чи через засоби електронної комунікації у вигляді текстового зображення.

Нині для пришвидшення оброблення інформації і введення її в облікові системи або формування власних цифрових сховищ і реєстрів, постала проблема оброблення таких документів. Традиційно такі документи оброблювались звичайними операторами персональних комп'ютерів, що є дуже трудомістким, дорогим та може містити певну кількість помилок через різноманітні фактори.

Для вирішення цієї проблеми існує спеціалізоване програмне забезпечення, яке завдяки використанню оптичного розпізнання символів надає інформацію про дані, що є на документі [1]. Такі системи можуть і не вилучати ключові значення з документів, але так чи інакше спрощують роботу операторів.

Як подальший розвиток таких систем було утворено системи вилучення даних, засновані на шаблонах та правилах [2-4]. Ці системи є досить надійними і якісно вилучають дані з документів, на які можна накласти раніше створені шаблони та правила. Недоліком їх є потреба вносити зміни в налаштування для оброблення документів зі зміненою структурою або взагалі впровадження нових налаштувань для нових типів документів. Мова документів, яку можуть розпізнати ці системи, залежить від використовуваного двигуна розпізнавання.

Подальшим кроком у розвитку систем вилучення інформації з цифрових документів були системи, що ґрунтувалися на нейронних мережах. Такі системи можуть швидше і коректніше вилучати дані [3], причому деякі системи навіть не потребують створення власних моделей для вилучення даних.

Мета статті — дослідити наявні системи аналізу зображень та вилучення даних, що засновані на нейронних мережах, та можливість їх застосування для оброблення україномовних документів.

ОСНОВНА ЧАСТИНА

Серед систем, заснованих на нейронних мережах, глибокому або машинному навчанні [3; 5] сьогодні найбільш відомі такі: Microsoft Azure Form Recognizer [6], Google Form Parser [7], Amazon Textracts [8] та Nanonets [9]. Перед проведенням дослідження коротко розглянемо інформацію, здобуту про ці системи:

Microsoft Azure Form Recognizer

Microsoft Azure Form Recognizer — це хмарна «когнітивна служба», що надає свої послуги через REST API, а отже, вона може працювати лише за наявності стабільного підімкнення до мережі Інтернет і потребує створення програмного забезпечення аби звичайні користувачі мали змогу користуватися нею. Система може працювати у таких режимах:

© К. О. Гордієнко, А. Б. Коба, Т. П. Довженко, 2020

• вбудовані моделі — доступні лише для англомовних чеків та візитних карток. Це попередньо навчені моделі, які не потребують додаткового навчання та налаштувань і вже готові до використання;

• нестандартні моделі — моделі, що були створені для певного набору документів за потребами зацікавлених осіб і не є доступними всім користувачам сервісу. Формуються такі моделі через веб-інтерфейс оператором, котрий виокремленням на зображенні необхідної для вилучення ділянки створює пари ключ–значення. Для навчання моделі потрібно мінімум п'ять зображень;

• автоматично створювані моделі через API — ці моделі, як і нестандартні, формуються з певного набору документів. Різниця в тому, що навчання відбувається в автоматичному режимі, без потреби залучення оператора. Через REST API надсилається набір однотипних документів (від п'яти), у відповідь система надсилає пари ключ-значення та координати їх місцезнаходження на зображенні.

За документацію Azure Form Recognizer може працювати з такими мовами: китайською (спрощена), датською, англійською (друкований та рукописний текст), французькою, німецькою, італійською, португальською та іспанською.

Google Form Parser

Google Form Parser — це частина хмарного комплексу розпізнавання документів Google Documents AI, а саму систему побудовано на технології машинного навчання. Функціонал працює тільки в разі підімкнення через мережу Інтернет, оброблення можливе лише через REST API.

Система підтримує розпізнавання текстів для документів із більш як 200 мовами, зокрема з українською.

Form Parser містить набір попередньо навчених моделей, переважно це податкові форми з США, але також можливе розпізнавання чеків та рахунків-фактур. Для випадків, коли необхідно розпізнати документи, для яких відсутня раніше створена модель, є можливість використовувати загальний розпізнавач, але в такому разі система просить надати мінімум як 10 однотипних документів для коректного вилучення.

Якщо використовується загальний розпізнавач, то система автоматично формує пари ключ-значення і вилучає таблиці з наданих зображень. Попереднє оброблення зображень необов'язкове, оскільки система може самостійно підібрати правильні налаштування.

Amazon Textracts

Amazon Textracts — це хмарна підсистема, що є частиною Amazon recognition API. Система автоматично формує пари ключ-значення та вилучає таблиці, а також може стати частиною оброблення за допомогою Amazon Augmented AI.

Хоча є базова можливість працювати через веб-інтрефейс, проте основна робота з підсистемою відбувається через REST API.

Згідно з документацією система може працювати лише з мовами, заснованими на латинському алфавіті та із символами Unicode.

Nanonets

Nanonets — це хмарний продукт, спрямований на вилучення даних із цифрових зображень, базуючись на глибокому навчанні.

Продукт дає можливість працювати як із попередньо навченими моделями (рахунки-фактури, чеки, ID-картки тощо), так і створювати модель для власного набору документів.

Для власних наборів документів система може автоматично вилучати лише табличні дані, для вилучення пар ключ-значення необхідно, щоб оператор у ручному режимі через веб-інтерфейси задав межу необхідних значень і прив'язав їх до «ключів». Для навчання моделі необхідно від 10 екземплярів однотипних документів.

Система має початкові можливості для коригування вилучених даних, але свої послуги надає переважно через REST API.

Оскільки однією зі складових частин системи є двигун оптичного розпізнавання символів Tesseract [9], то продукт підтримує більш як 240 мов, зокрема українську.

Результати дослідження

Для перевірки працездатності цих систем з україномовними документами було створено набір документів зі структурою найпростіших рахунків-фактур, в яких застосовується лише українська мова. Усі сформовані документи завантажувались у раніше описувані системи, і за необхідності проводилось створення моделей для цих документів.

Дослідження відбувалося навіть на тих системах, які за документацією не підтримують українську мову. До таких систем належать Microsoft Azure Form Recognizer та Amazon Textract. Ці системи вилучають дані менш ніж за хвилину, але через відсутність підтримання української мови, для вилучення даних використовують латинский алфавіт. Отже, замість україномовних значень системи вилучають набір максимально схожих латинських символів.

Google Form Parser успішно вилучає україномовні значення, але для ключів вилучає імена з помилками, часто застосовуючи символи латинського алфавіту. Проте варто зазначити, що перевагою системи є можливість вилучення даних навіть із документів, які мають певні нахили зображення.

Nanonets успішно вилучає україномовні дані на документах без викривлень. Через те що оператор самостійно створює ключі, до яких прив'язується значення, проблеми як з Google Form Parser немає. Nanonets їх просто не формує. Якщо на зображенні є істотні викривлення, система не здатна правильно співвіднести дані з ключами, через що вилучається вся знайдена інформація в одне поле (за аналогією роботи звичайних OCR-систем).

Таким чином, для оброблення україномовних документів більш за все підходить Google Form Parser, оскільки не потребує попереднього навчання моделей і має здатність до попереднього оброблення зображень.

ВИСНОВКИ

Сьогодні немає систем, котрі можуть самостійно визначати, саме які дані необхідні для вилучення з україномовних документів. Наявні системи потребують створення програмного забезпечення, які відіграватимуть роль обкладинки для функціонала систем, що передають свою інформацію через REST API.

Навіть якщо взяти систему, яка показала себе найкраще (у даному дослідженні це Google Form Parser), необхідність постійного підімкнення до мережі Інтернет може стати серйозною перепоною для використання такого продукту в певних сферах діяльності.

Список використаної літератури

1. ***Lebourgeois F., Henry J.-L., Emptoz H.*** *An OCR System for Printed Documents. 1992. P. 83–86.*
2. ***Sudharshan Chandra Babu*** *from Nanonets (2020). Automating Receipt Digitization with OCR and Deep Learning [Електронний ресурс]. URL: https://nanonets.com/blog/receipt-ocr/*
3. ***Семенов С.*** *Как научить машину понимать инвойсы и извлекать из них данные [Електронний ресурс]. URL: https://habr.com/ru/company/abbyy/blog/440310/*
4. ***Intellix*** *– End-User Trained Information Extraction for Document Archiving / D. Schuster, K. Muthmann, D. Esser [et al.] // Proceedings of the International Conference on Document Analysis and Recognition, ICDAR. 10.1109/ICDAR.2013.28.*
5. ***Azure*** *vs AWS vs GCP (Part 2: Form Recognizers) [Електронний ресурс]. URL: https://cazton.com/blogs/executive/form-recognition-azure-aws-gcp.*
6. ***Form*** *Recognizer documentation [Електронний ресурс]. URL: https://docs.microsoft.com/en-us/azure/cognitive-services/form-recognizer*
7. ***Document*** *AI Documentation [Електронний ресурс]. URL: https://cloud.google.com/document-ai/docs*
8. ***Amazon*** *Textract Developer Guide [Електронний ресурс]. URL: https://docs.aws.amazon.com/textract/latest/dg/what-is.html*
9. ***Nanonets*** *[Електронний ресурс]. URL: https://nanonets.com/*

К. А. Гордиенко, А. Б. Коба, Т. П. Довженко

ИССЛЕДОВАНИЕ СИСТЕМ РАСПОЗНАВАНИЯ ТЕКСТА И УДАЛЕНИЕ ДАННЫХ ДЛЯ УКРАИНОЯЗЫЧНЫХ ДОКУМЕНТОВ

Рассмотрено существующее программное обеспечение, основной задачей которого является извлечение информации из оцифрованных документов. Из всего программного обеспечения отбиралось то, что основывается на технологиях нейронных сетей и глубокого обучения. Для извлечения информации из документов может использоваться ручной труд операторов персональных компьютеров, требующий много времени и не исключающий влияние человеческого фактора, а также оцифровка документов с последующей обработкой в программном обеспечении, основанной на принципе подчинения документов шаблонам и правилам, что может влиять на скорость обработки данных и необходимость вносить изменения в настройки через изменение типа документа.

В статье ставится задача исследовать существующее программное обеспечение для извлечения данных из цифровых документов, основанное на технологии нейронных сетей, и их применимость к украиноязычным документам. Для этого был создан простой набор счетов-фактур, которые загружались в систему.

Разработка системы для извлечения информации из оцифрованных украиноязычных документов с помощью нейронных сетей ускорит обработку данных, предоставит возможность для их обработки в зависимости от сферы деятельности пользователя этого программного обеспечения.

Установлено, что в настоящее время нет систем, которые могут самостоятельно определять, какие данные необходимы для извлечения из украиноязычных документов. Имеющиеся системы требуют создания программного обеспечения, которые будут играть роль обложки для функционала систем, которые передают свою информацию через REST API. Обосновано, что лучшей системой является Google Form Parser, однако она требует постоянного подключения к сети Интернет, что может стать серьезным препятствием для использования такого продукта в определенных сферах деятельности.

Ключевые слова: оптическое распознавание символов; нейронные сети; глубокое обучение; машинное обучение; извлечение данных.

K. O. Hordienko, A. B. Koba, T. P. Dovzhenko

RESEARCH OF TEXT RECOGNITION SYSTEMS AND DATA REMOVAL FOR UKRAINIAN-LANGUAGE DOCUMENTS

This article discusses the existing software, the main task of which is to extract information from digitized documents. From all the software was selected what is based on neural network technology and deep learning. To extract information from documents, manual work of personal computer operators can be used, which takes a long time and does not exclude the influence of the human factor, as well as digitization of documents with further processing in software based on the principle of subordination of documents to templates and rules, data processing speed and the need to make changes to the settings due to a change in the type of document.

The article aims to investigate the existing software for extracting data from digital documents based on neural network technology, and their applicability to Ukrainian-language documents. To do this, a simple set of invoices was created and uploaded to the system.

The development of a system for extracting information from digitized Ukrainian-language documents using neural networks will speed up data processing, provide an opportunity for their processing depending on the scope of the user of this software.

It is established that at present, there are no systems that can independently determine what data is needed for extraction from Ukrainian-language documents. Existing systems require the creation of software that will act as a cover for the functionality of systems that transmit their information through the REST API. Google Form Parser is considered to be the best system, but it requires a constant connection to the Internet, which can be a serious obstacle to the use of such a product in certain areas of activity.

Keywords: optical character recognition; neural network; deep learning; machine learning; data extraction.