Issue 8, 2020

https://doi.org/10.23939/sisn2020.08.001

УДК 004.421.6

МЕТОД ПЕРВИННОЇ ОБРОБКИ СЛАБОСТРУКТУРОВАНИХ МЕДИЧНИХ ДАНИХ

Дмитро Бичко[1], Віра Шендрик[2], Юлія Парфененко[3]

[1, 2, 3] Сумський державний університет
[1] d.bychko11@gmail.com, ORCID: 0000-0002-6854-945X
[2] v.shendryk@cs.sumdu.edu.ua, ORCID: 0000-0001-8325-3115
[3] yuliya_p@cs.sumdu.edu.ua, ORCID: 0000-0003-4377-5132

У статті розглянуто підхід до первинної обробки слабоструктурованих текстових даних медичних протоколів, що зберігаються та розповсюджуються у вигляді файлів у pdf-форматі. Актуальність цієї роботи зумовлена відсутністю універсальної структури подання медичних протоколів та методів їхньої обробки. У ході роботи вирішено задачу первинної обробки даних клінічних протоколів на прикладі уніфікованого клінічного протоколу первинної, вторинної (спеціалізованої) та третинної (високоспеціалізованої) медичної допомоги. Розроблено метод первинної обробки даних для створення чіткої структури симптомів хвороби. Першим етапом структуризації даних клінічного протоколу запропоновано розділення інформації з протоколу на чотири базові частини, що дозволяє пришвидшити його конвертування в інші формати. Цей процес реалізовано за допомогою алгоритму, який розроблено мовою програмування C#. Запропонований алгоритм реалізує парсинг інформації з файлу, що представлений у pdf-форматі, та перетворює її у файл txt. Після цього виконується обробка одержаної інформації, що полягає у синтаксичному аналізі тексту протоколу та виділенні структурних частин протоколу, що відповідають заголовкам розділів: титульний аркуш; вступ; перелік скорочень, що використовуються у протоколі; основна частина протоколу; перелік літературних джерел. Назву хвороби у медичному протоколі ідентифікують, порівнюючи дані з протоколу та переліком назв захворювань, що представлені у світовій класифікації МКХ-10. Було проаналізовано заголовки "Вступ", "Перелік скорочень, що використовуються у протоколі" та основної частини протоколу і запропоновано алгоритм видалення малоінформативних розділів з початку протоколу, наприклад, літературних джерел. Також розроблено алгоритм пошуку інформації в основній частині медичного протоколу шляхом обробки вхідних даних за таблицями, схемами, заголовками, словами, фразами та спеціальними символами. У результаті роботи алгоритмів обробки клінічного протоколу формується новий файл клінічного протоколу, що має приблизно втричі менший обсяг порівняно з початковим файлом. Він містить у собі лише змістовну інформацію з клінічних протоколів, що прискорить подальшу роботу з цим файлом, а саме його використання в системі підтримки прийняття медичних рішень. Представлено картку хвороби на основі медичного протоколу у форматі JSON.

Ключові слова: слабоструктуровані медичні дані, клінічний протокол, первинна обробка, природомовні тексти, метод, псевдокод.

Вступ

У наш час людство накопичило великий обсяг медичних даних. Наявні інформаційні технології здатні поліпшити процес постановки діагнозу та підвищити якість сучасних медичних послуг. У своїй практиці лікарі використовують клінічні протоколи лікування хвороб, які містять короткий

опис хвороби, методи діагностики та лікування. Процес прийняття відповідного рішення лікарем повинен спиратися на інформацію, що представлена у клінічних протоколах первинної медичної діагностики та лікування хвороб, які являють собою набір слабоструктурованих природомовних текстових даних. Це ускладнює пошук лікарем необхідної інформації для постановки діагнозу та визначення заходів надання медичної допомоги. Тому для підвищення оперативності прийняття медичних рішень під час діагностики та лікування необхідно реалізувати програмні засоби видобування медичних знань та підтримки прийняття рішень. Для цього необхідно розробити структуру подання даних з медичних протоколів та реалізувати алгоритми їхньої обробки. Першочерговою задачею, яку необхідно вирішити, є первинна обробка тестів медичних протоколів з метою виділення з протоколу змістовної інформації та її перетворення у формат, що забезпечить її автоматичне опрацювання у прийнятті медичних рішень.

Постановка проблеми

У цій статті розглянуто підхід до первинної обробки слабоструктурованих текстових даних медичних протоколів, що зберігаються та розповсюджуються у вигляді файлів у pdf-форматі. Інформацію, наведену у клінічних протоколах, поділено на кілька змістовних розділів, але немає єдиної форми її представлення через різний обсяг та формат інформації, наявність схем та рисунків, що значно ускладнює процес обробки протоколу, перегляду та швидкого пошуку необхідних даних. На сьогодні систематизація даних медичних протоколів реалізована на рівні каталогізації файлів з повними текстами медичних протоколів за назвою хвороби чи групи захворювань.

Для автоматичного видобування знань з клінічних протоколів лікування хвороб сьогодні використовують спрощені пошукові методи, наприклад, пошук за словом чи номером протоколу, які не враховують ані контекст запитів, ані семантичні складові, що призводить до повільного пошуку та отримання недостовірних результатів. Через це потреби користувачів при пошуку необхідної інформації з клінічних протоколів задовольняються не повною мірою. Проблема видобування, обробки, перетворення та зберігання неструктурованих та слабоструктурованих даних медичних протоколів для швидшої взаємодії користувача з інформаційною системою є досить актуальною. Тому необхідно структурувати дані медичних протоколів та подати у чітко визначеному форматі, що дозволить швидко оброблювати вхідну інформацію та точніше діагностувати хворобу, спираючись на вже існуючі експертні знання, що знижує вплив людського чинника на прийняття медичних рішень.

Статтю присвячено розробці методу попередньої обробки слабоструктурованих текстів медичних протоколів, що має на меті виділити інформацію за основними критеріями пошуку, тобто скоротити pdf-файл, сформувати json із попередньо оброблених pdf-файлів, тобто виконати перетворення файлів із медичними протоколами у таку форму, яка дозволяє якісно порівнювати дані з медичного протоколу із базою даних наявних діагнозів та симптомів.

Аналіз останніх досліджень та публікацій

Сучасні інформаційні технології дають можливість автоматично опрацьовувати тексти, використовуючи методи інтелектуального аналізу даних, штучного інтелекту та комп’ютерної лінгвістики. Основними принципами, які значно покращують пошук інформації є:

- використання структурованих даних для пошуку;
- зосередженість на основній меті, а саме задоволення потреб користувачів;
- фокус на пошуку інформації, а не даних.

Більшість медичних даних представлені у слабоструктурованій або неструктурованій формі. Порівняно зі структурованими даними, вони є природнішими та легшими для сприйняття. На цей час представлені різні методології опрацювання тексту: виділення відношень між поняттями, сутностями та атрибутами. Але через великі об’єми інформації виникає проблема встановлення цих відношень через відсутність уніфікованого алгоритму. У роботі [1] представлено підхід у розробленні декількох модулів, що дозволяють робити синтаксичний (токенізація, тегування частин мови) та семантичний (розпізнавання різних частин мови) аналіз тексту. Кращим підходом є

написання специфічних парсерів залежно від типу документів [2, 3]: рентгенівських звітів, виписок лікарів, медичної документації тощо.

Для обробки даних шляхом парсингу даних використовують таке програмне забезпечення: cTAKES, MedTagger, MetaMap, KnowledgeMap та ін. [4, 5, 6]. Вирішити задачі структурування та обробки інформації можна за допомогою Office Open XML з використанням Health Level 7 Clinical Document Architecture, Info Path [7], XML Medical Markup Language [8], які широко використовують у сучасній медицині. Але їхні можливості доволі обмежені, і тому вони не можуть бути використані для опрацювання даних із клінічних протоколів через відсутності розширення користувальницького функціоналу та ймовірності створення власних методів обробки даних, які розбігаються з запропонованими [8].

JSON формат використовують для подання медичних даних в інформаційних системах [7], забезпечуючи швидку взаємодію між різними видами приладів (мобільні телефони, персональні комп'ютери, планшети), а також інтеграцію між системами прийняття рішень, які мають різну архітектуру.

Аналіз попередніх досліджень показав, що для медичних протоколів ще не розроблена універсальна структура подання даних та методи їхньої обробки для роботи з українською мовою. Для чіткої структуризації симптомів хвороби на основі текстів клінічних протоколів необхідною є попередня обробка даних.

Формулювання мети статті

Під час первинного огляду пацієнта необхідно зібрати низку симптомів та швидко їх опрацювати. На цьому етапі необхідно чітко виділити всі явні та неявні проблеми пацієнта та поставити діагноз, тобто прийняти відповідне рішення із зіставлення можливих ознак хвороби з інформацією про існуючі хвороби. При постановці діагнозу лікар використовує власні знання та знання, формалізовані у вигляді медичних протоколів. Але через велику кількість схожих на перший погляд симптомів хвороби лікареві досить важко визначити правильний діагноз. Тому розробка механізму первинної обробки великих об'ємів слабоструктурованих даних з медичних протоколів з метою їх зберігання та подальшого використання в системі підтримки прийняття медичних рішень дозволить підвищити якість лікування хворих та знизити вірогідність визначення помилкового діагнозу.

Виклад основного матеріалу

Медичні протоколи зберігаються та розповсюджуються в електронному вигляді у pdf-форматі, що дає можливість знайти в них необхідну інформацію шляхом пошуку за окремими словами чи словосполученнями. Але через велику кількість сторінок протоколів (у середньому – понад 50), наявність однакових клінічних симптомів у різних протоколах виникає потреба у значній оптимізації організації інформації з клінічних протоколів для зручнішої взаємодії з нею.

Розглянемо задачу обробки даних клінічних протоколів на прикладі уніфікованого клінічного протоколу первинної, вторинної (спеціалізованої) та третинної (високоспеціалізованої) медичної допомоги, оскільки цей тип протоколів містить усі необхідні дані, щоб сформувати структуру для побудови інформаційної системи, яка призначена для аналізу вхідних симптомів від хворого та визначення ймовірного захворювання. Цей процес охоплює фрагментацію речень, синтаксичний та морфологічний аналіз. Пропонуємо поділити уніфікований клінічний протокол на чотири базові частини (рис. 1) та розділити його на чотири окремі файли у pdf-форматі для швидшого конвертування файлів у інші формати.

На першому етапі потрібно перетворити файли з pdf-формату у зручнішу для обробки форму. Для цього розроблено алгоритм, представлений на рис. 3, який дасть змогу парсити інформацію у існуючому форматі та перетворювати її у txt-файл. Цей алгоритм є модифікацією алгоритму, запропонованого в роботі [9], що полягає у реалізації функції збереження отриманого файлу після того, як буде згенерований рядок у програмі. Алгоритм реалізований за допомогою мови програмування C# [9] та запускається через створений після компіляції exe-файл у середовищі розробки програмного забезпечення Microsoft Visual Studio 2013.

Рис. 1. Структура уніфікованого клінічного протоколу

Обробка уніфікованого клінічного протоколу здійснюється у кілька етапів, послідовність яких показано на рис. 2.

Рис. 2. Схема обробки уніфікованого клінічного протоколу

Суть алгоритму (рис. 3) полягає в обробці вхідних даних, які представлені файлами у форматі pdf. Існуючий файл додається у папку з exe-файлом, потім потрібно запустити програму. Вона підключає можливості двох бібліотек iTextSharp.text.pdf та iTextSharp.text.pdf.parser для взаємодії з парсингом та обробкою тексту у pdf-форматі. Далі відбувається завантаження файлу та створення об'єкта з необхідною вхідною інформацією. Він містить у собі весь документ. За допомогою бібліотеки виконують обробку об'єкта шляхом перетворення кожної сторінки у рядок за допомогою циклу. У результаті отримуємо рядок, що містить усі сторінки вхідного документу. Цей процес повторюється для всіх частин протоколу. Після обробки всіх чотирьох файлів, які створені у результаті виконання першого етапу, що містять частини протоколу, отримуємо один файл, який представлено на рис. 4. Він містить усю інформацію у txt-форматі. Далі зберігаємо його у txt-форматі та відображаємо у папці з exe-файлом.

Далі необхідно обробити отриману інформацію. Для цього треба проаналізувати текст протоколу шляхом пошуку структурних заголовків, що є заголовками розділів. У результаті отримуємо окремі пункти:

1. Титульний аркуш.
2. Вступ.
3. Перелік скорочень, що використовуються у протоколі.
4. Основну частину протоколу.
5. Перелік літературних джерел.

Рис. 3. Алгоритм роботи парсеру даних з медичних протоколів

Рис. 4. Консоль програми та згенерований файл у txt-форматі

Із цих пунктів треба обрати ті, з яких є можливість видобути змістовну інформацію:

- з першого пункту потрібно взяти назву хвороби;
- другий і третій пункт необхідно пропустити, оскільки вони містять загальні фрази описового характеру;
- четвертий та п'ятий пункти потрібно аналізувати, виокремлюючи основні закономірності запису тексту, структури заголовків та шуканих слів;
- шостий пункт містить список літератури, тому його слід пропустити.

Розглянемо метод обробки тексту уніфікованого клінічного протоколу детальніше. На титульному аркушу виділимо назву хвороби. Алгоритм цього процесу представлено у вигляді псевдокоду на рис. 5. Назви хвороб у процесі їхньої ідентифікації порівнюються з переліком хвороб за світовою класифікацією хвороб МКХ-10. Це дає змогу перевіряти коректність представлених назв у відповідних протоколах та еталонному довіднику (МКХ-10).

```
Початок
Почати пошук, поки не знайдемо вперше значення "20%%", де % – будь-яке числове значення
Збережемо данні у змінну, які знаходяться до "20%%"
Порівняємо отримані дані з масивом назв хвороб МКХ-10
Якщо знайшли збіг
    То зберігаємо у змінну
    Завершуємо пошук
    Кінець
Інакше продовжити пошук, перейшовши на крок 4
```

Рис. 5. Псевдокод алгоритму пошуку назви хвороби

Наступним кроком відбувається аналіз заголовків “Вступ”, “Перелік скорочень, що використовуються у протоколі” та заголовки з основної частини протоколу. Як видно з рис. 1, перші два заголовки мають формат “перше слово з великої літери, наступні – з маленької”, а заголовок основної частини – “римська цифра та текст з усіма великими літерами”. Алгоритм процесу пошуку заголовків основної частини представлений на рис. 6.

```
Початок
Почати пошук, поки не знайдемо вперше "1" або "I"
Якщо знайшли
    То почати пошук, поки не знайдемо вперше значення "[А-Я][А-Я]", де [А-Я] – будь-яка велика літера від А до Я
    Видалити весь текст до "1 [А-Я][А-Я]" або "I [А-Я][А-Я]"
    Кінець
Інакше перейти до кроку 2
```

Рис. 6. Псевдокод алгоритму пошуку заголовку, що містить на початку “1” або “I” та текст з усіма великими літерами

Після пошуку та обробки заголовків виконується перехід до обробки переліку літературних джерел у кінці протоколу. Для цього використано алгоритм, представлений у псевдокоді на рис. 7. Він дозволяє знайти у кінці документа літературні джерела та видалити їх, що зменшить обсяги кінцевого вихідного файлу після його обробки.

```
Початок
Переходимо в кінець документу
Почати пошук з кінця документу у початок поки не знайдемо вперше слово "літерат%"
Якщо знайшли
    То почати пошук вгору, поки не знайдемо подвійний відступ
    Видалити весь текст починаючи з подвійного відступу до кінця документу
    Кінець
Інакше перейти до кроку 2
```

Рис. 7. Псевдокод алгоритму пошуку переліку джерел

У результаті виконання всіх зазначених дій (виділення назви хвороби, відсікання початку та кінця документу (рис. 5–7) отримуємо протокол, який складається з 4 частин, що є скороченою версію уніфікованого медичного протоколу з визначеною для подальшої обробки змістовною інформацією, що включає у себе симптоми та значення їхніх параметрів.

У результаті обробки отримуємо pdf-файл, який містить приблизно втричі менше інформації (наприклад, файл обсягом 429 кілобайтів зменшився до 138 кілобайтів), ніж було до його обробки. Він містить у собі один медичний протокол, в якому залишилася лише змістовна інформація.

Сформований pdf-файл з обробленим клінічним протоколом містить менший об’єм даних і зберігає необхідну інформацію у вигляді необхідних для роботи системи підтримки прийняття медичних рішень фраз, заголовків, слів, схем, таблиць та спеціальних Unicode символів.

Після обробки pdf-файлу наступним етапом відбувається генерація JSON-файлу для створення картки хвороби, яка містить поле зі значенням назви хвороби, перелік назв симптомів у вигляді списку та їхніх фактичних значень.

Видобування даних для подальшого їх перетворення у JSON-формат проводиться за напрямами (рис. 8):

1. за таблицями;
2. за схемами;
3. за заголовком;
4. за словом;
5. за фразою;
6. за спеціальними Unicode символами.

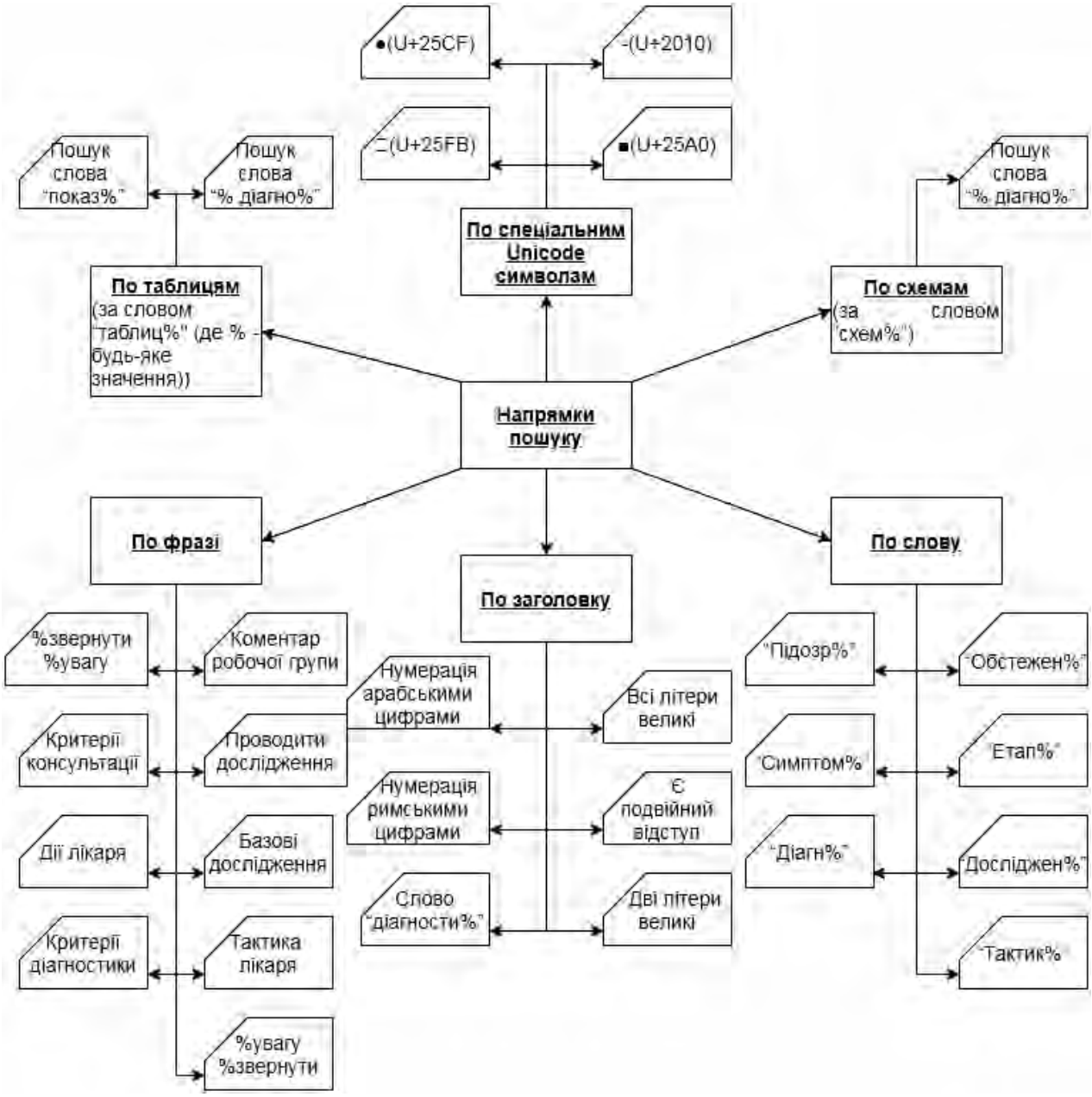

Рис. 8. Напрямки пошуку в основній частині протоколу

Основні напрямки та правила пошуку були розроблені у ході ґрунтовного аналізу структури документа (рис. 1), що зробить пошук необхідної інформації якіснішим порівняно з вже існуючими розробками.

Прикладом картки хвороби на основі медичного протоколу у форматі JSON є:

```
{
  “name”:”Епідемічний паротит”,
  “icd10”:”B26”,
  “symptom1”:”температура до 40°C”,
  “symptom2”:”висока температура 5-7 днів”,
  “symptom3”:”біль в області вух”,
  “symptom4”:”больовий симптом при жуванні чи ковтанні”,
  “symptom5”:”підвищене слиновиділення”,
  “symptom6”:”різкий біль при вживанні кислої їжі,
  “symptom7”:”збільшення щік”,
  “symptom8”:”припухлість зони перед вушною раковиною, що збільшується в розмірах”,
  “symptom9”:”мочка вуха відстовбурчується вперед і догори”,
  “symptom10”:”біль при пальпації слинних залоз”
}
```

де name – назва хвороби, icd10 – код за МКХ-10, symptom – симптом.

Після створення окремого JSON-файла з карткою хвороби відбудеться його збереження до документоорієнтованої бази даних. Це дає змогу структурувати та зберігати дані з клінічних протоколів та у подальшому пришвидшити пошук у системі за вхідними даними.

Висновки

У статті досліджено проблематику щодо обробки та зберігання слабоструктурованої інформації. Запропоновано метод первинної обробки слабоструктурованих медичних даних на прикладі уніфікованого медичного протоколу. Розроблено парсер для слабоструктурованих медичних протоколів. У результаті виділено інформацію за основними критеріями пошуку та підготовлено файли з медичними протоколами у такій формі, яка дозволить якісно порівнювати дані з медичного протоколу із базою існуючих симптомів для створення структурованого набору даних, який описує хворобу. У подальших дослідженнях буде розроблено алгоритм пошуку даних, які представлені у медичних протоколах у вигляді симптомів та їхніх значень (наприклад, наявність температури понад 37 градусів є показником певного захворювання). У результаті, програмна реалізація згаданого алгоритму за допомогою однієї з мов програмування (C++, C# або Java) дозволить обробляти вхідні дані, які отримані у результаті цієї роботи та представлені у більш структурованому вигляді файлів. Після цього процесу отримано набір даних у вигляді: назва хвороби, симптом та його значення (кількісне або якісне) та сформовано картки хвороб, що містять у собі назву хвороби, код за МКХ-10, симптом 1 та його значення, симптом 2 та його значення та всі інші симптоми, які характеризують дане захворювання у форматі JSON. Потім ця інформація буде занесена до бази даних для структурованого зберігання. Наступним кроком буде розроблено метод швидкої обробки вхідних запитів лікаря через інтерфейс інформаційної системи, а саме обрання симптомів та їхніх показників. Після реалізації методу обробки симптомів необхідно вирішити завдання визначення найімовірнішої хвороби, маючи її характеристики, отримані у ході структуризації даних з медичних протоколів. У результаті буде створено інформаційну систему підтримки прийняття рішень, що дозволить лікареві точніше ставити діагноз.

Список літератури

1. Jensen, K., Soguero-Ruiz, C., Oyvind Mikalsen, K., Lindsetmo, R., Kouskoumvekaki, I., Girolami, M., Augestad, K. M. (2017). Analysis of free text in electronic health records for identification of cancer patient trajectories. *Scientific Reports*, *7*(1). doi:10.1038/srep46226

2. Kung, R., Ma, A., Dever, J. B., Vadivelu, J., Cherk, E., Koola, J. D., Ho, S. B. (2015). Mo1043 a natural language processing Alogrithm for identification of patients with cirrhosis from electronic medical records. *Gastroenterology*, *148*(4), S-1071–S-1072. doi:10.1016/s0016-5085(15)33662-3

3. Li, D., Azoulay, P., & Sampat, B. N. (2017). The applied value of public investments in biomedical research. *Science*, *356*(6333), 78–81. doi:10.1126/science.aal0010

4. Patel, R., Lloyd, T., Jackson, R., Ball, M., Shetty, H., Broadbent, M., Taylor, M. (2015). Mood instability is a common feature of mental health disorders and is associated with poor clinical outcomes. *BMJ Open*, *5*(5), e007504–e007504. doi:10.1136/bmjopen-2014-007504

5. Wi, C., Sohn, S., Rolfes, M. C., Seabright, A., Ryu, E., Voge, G., Juhn, Y. J. (2017). Application of a natural language processing algorithm to asthma ascertainment. An automated chart review. *American Journal of Respiratory and Critical Care Medicine*, *196*(4), 430–437. doi:10.1164/rccm.201610-2006oc

6. Afzal, N., Sohn, S., Abram, S., Scott, C. G., Chaudhry, R., Liu, H., Arruda-Olson, A. M. (2017). Mining peripheral arterial disease cases from narrative clinical notes using natural language processing. *Journal of Vascular Surgery*, *65*(6), 1753–1761. doi:10.1016/j.jvs.2016.11.031

7. O365devx. (n.d.). Working with XML Schemas in InfoPath. Technical documentation, API, and code examples | Microsoft Docs. https://docs.microsoft.com/en-us/office/client-developer/infopath/form-templates/ working-with-xml-schemas-in-infopath

8. The Latest MML (Medical Markup Language) Version 2.3 – XML-Based Standard for Medical Data Exchange/Storage. (n.d.). ResearchGate. https://www.researchgate.net/publication/10675074_The_Latest_MML_Medical_Markup_Language_Version_23_-_XML-Based_Standard_for_Medical_Data_ExchangeStorage

9. Parsing PDF Files using iTextSharp (C#, .NET). (n.d.). Square PDF .NET. https://www.squarepdf.net/parsing-pdf-files-using-itextsharp

References

1. Jensen, K., Soguero-Ruiz, C., Oyvind Mikalsen, K., Lindsetmo, R., Kouskoumvekaki, I., Girolami, M., Augestad, K. M. (2017). Analysis of free text in electronic health records for identification of cancer patient trajectories. *Scientific Reports*, *7*(1). doi:10.1038/srep46226

2. Kung, R., Ma, A., Dever, J. B., Vadivelu, J., Cherk, E., Koola, J. D., Ho, S. B. (2015). Mo1043 a natural language processing Alogrithm for identification of patients with cirrhosis from electronic medical records. *Gastroenterology*, *148*(4), S-1071-S-1072. doi:10.1016/s0016-5085(15)33662-3

3. Li, D., Azoulay, P., & Sampat, B. N. (2017). The applied value of public investments in biomedical research. *Science*, *356*(6333), 78-81. doi:10.1126/science.aal0010

4. Patel, R., Lloyd, T., Jackson, R., Ball, M., Shetty, H., Broadbent, M., Taylor, M. (2015). Mood instability is a common feature of mental health disorders and is associated with poor clinical outcomes. *BMJ Open*, *5*(5), e007504–e007504. doi:10.1136/bmjopen-2014-007504

5. Wi, C., Sohn, S., Rolfes, M. C., Seabright, A., Ryu, E., Voge, G., Juhn, Y. J. (2017). Application of a natural language processing algorithm to asthma ascertainment. An automated chart review. *American Journal of Respiratory and Critical Care Medicine*, *196*(4), 430–437. doi:10.1164/rccm.201610-2006oc

6. Afzal, N., Sohn, S., Abram, S., Scott, C. G., Chaudhry, R., Liu, H., Arruda-Olson, A. M. (2017). Mining peripheral arterial disease cases from narrative clinical notes using natural language processing. *Journal of Vascular Surgery*, *65*(6), 1753–1761. doi:10.1016/j.jvs.2016.11.031

7. O365devx. (n.d.). Working with XML Schemas in InfoPath. Technical documentation, API, and code examples | Microsoft Docs. https://docs.microsoft.com/en-us/office/client-developer/infopath/form-templates/ working-with-xml-schemas-in-infopath

8. The Latest MML (Medical Markup Language) Version 2.3 – XML-Based Standard for Medical Data Exchange/Storage. (n.d.). ResearchGate. https://www.researchgate.net/publication/10675074_The_Latest_MML_Medical_Markup_Language_Version_23_-_XML-Based_Standard_for_Medical_Data_ExchangeStorage

9. Parsing PDF Files using iTextSharp (C#, .NET). (n.d.). Square PDF .NET. https://www.squarepdf.net/parsing-pdf-files-using-itextsharp

THE METHOD OF PRIMARY PROCESSING OF POORLY STRUCTURED MEDICAL DATA

Dmytro Bychko[1], Vira Shendryk[2], Yuliia Parfenenko[3]

[1, 2, 3] Sumy State University,
[1] d.bychko11@gmail.com, ORCID: 0000-0002-6854-945X
[2] v.shendryk@cs.sumdu.edu.ua, ORCID: 0000-0001-8325-3115
[3] yuliya_p@cs.sumdu.edu.ua, ORCID: 0000-0003-4377-5132

The article deals with the approach to the primary processing of poorly structured medical protocol textual data stored and disseminated as pdf files. The relevance of this work is due to the lack of a universal structure for the presentation of medical protocols and methods of their processing. In the course of the work, the problem of primary processing of clinical protocol data was solved by the example of a unified clinical protocol of primary, secondary (specialized) and tertiary (highly specialized) medical care. The method of primary data processing was developed to create a clear structure of the symptoms of the disease. The first step in structuring clinical protocol data is to divide the protocol information into four basic parts, which allows it to be quickly converted to other formats. This process is implemented using an algorithm developed in C # programming language. The proposed algorithm parses the information from a pdf file and converts it to a txt file. After that, the received information is processed, which consists in the syntactic analysis of the text of the protocol and selection of the structural parts of the protocol corresponding to the headings of the sections: title page; introduction; a list of abbreviations used in the protocol; the main part of the protocol; list of literary sources. The identification of the disease name in the medical protocol is performed by comparing the protocol data and the list of disease names, presented in the world classification MKH-10. The headings "Introduction", "List of abbreviations used in the protocol" and the main part of the protocol were analyzed and the algorithm for removing uninformed sections from the beginning of the protocol, for example, literature sources, was proposed. An algorithm for finding information in the main part of the medical protocol by processing input data by: tables, diagrams, headings, words, phrases and special symbols are also proposed. As a result of the clinical protocol processing algorithms, a new clinical protocol file is generated, which is three times smaller than the original file. It contains only meaningful information from clinical protocols that will speed up further work on this file, namely its use in medical decision support. The disease card based on a medical protocol in JSON format is presented.

Key words: poorly structured medical data, clinical protocol, primary processing, naturalistic texts, method, pseudocode