МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ЧОРНОМОРСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ПЕТРА МОГИЛИ

Мельничук Іван Олегович

УДК 004.43

Інтелектуальна система on-line перекладу на базі хмарних сервісів

122 – Комп’ютерні науки

Автореферат
магістерської кваліфікаційної роботи на здобуття освітньої кваліфікації
«Магістр комп’ютерних наук»

Миколаїв – 2021

Магістерська кваліфікаційна робота є рукопис.

Робота виконана в Чорноморському національному університеті імені Петра Могили Міністерства освіти і науки України на кафедрі інтелектуальних інформаційних систем

Науковий керівник: д.т.н., професор, професор кафедри інтелектуальних інформаційних систем Гожий Олександр Петрович

Рецензент: к.ф.м.н., доцент, доцент кафедри комп'ютерної інженерії Пузирьов Сергій Володимирович

Захист відбудеться «22» лютого 2021 р. о 9^{30} год. на засіданні екзаменаційної комісії (ауд. 2-403) у Чорноморському національному університеті імені Петра Могили за адресою: 54003, м. Миколаїв, вул. 68-ми Десантників, 10.

З магістерською кваліфікаційною роботою можна ознайомитися в бібліотеці Чорноморського національного університету імені Петра Могили за адресою: 54003, м. Миколаїв, вул. 68-ми Десантників, 10.

Автореферат представлений «16» лютого 2021 р.

Секретар
екзаменаційної комісії,
к.пед.н., доцент Н. М. Болюбаш

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми магістерської роботи полягає в дослідженні і застосуванні ефективних можливостей хмарних когнітивних систем для створення on-line систем перекладу.

Метою магістерської кваліфікаційної роботи є дослідження і застосування ефективних можливостей хмарних когнітивних систем для створення on-line систем перекладу.

Об'єктом дослідження є процеси обробки природньої мови в хмарних сервісах.

Предметом є методи та сервіси обробки природньої мови в хмарних сервісах.

Практичне значення даної магістерської кваліфікаційної роботи полягає у можливості застосування методів та сервісів обробки природньої мови для стоврення on-line систем перекладу.

Результати даної магістерської кваліфікаційної роботи було надруковано у тезах XXIII Всеукраїнської науково-методичної конференції «Могилянські читання – 2021» у секції Комп'ютерні науки.

Магістерська кваліфікаційна робота складається із вступу, 6 розділів, висновків, додатків. Загальний обсяг роботи складає 100 сторінок, 22 рисунків, 8 таблиць та 23 посилання на літературні джерела.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У вступі магістерської кваліфікаційної роботи обґрунтовано актуальність обраної теми, сформульовано мету і задачі дослідження, визначено предмет та об’єкт дослідження.

У першому розділі наведено огляд проблем обробки та структуризації людського мовлення. Розглянуто обробку природньої мови та єтапи її історичного розвитку.

Аналіз людського мовлення ускладнюєть синтаксична, смислова, відмінкова та референційна неоднозначності, а також проблеми синонімії. Розв'язання таких типів неоднозначностей можливе за допомогою введення додаткових значень, які збільшать знання програми про ту чи іншу галузь. Сьогодні програм, здатних «розуміть» усі типи неоднозначностей у великому спектрі галузей, не існує, проте є програми, що можуть коректно реагувати на неоднозначності у дуже вузьких сферах. До них входять системи:

- Видобування даних: вивчення даних, пошук зв'язків та закономірностей між ними
- Синтез мовлення: озвучення/прочитання тексту голосом, який є наближеним до природного.
- Розпізнавання мови: розпізнавання тексту з картинок та аудио файлів.
- Генерування природної мови: конвертування комп'ютерних даних у природну мову людини.
- Машинний переклад: автоматичний переклад з однієї людської мови на іншу.
- Питально-відповідальні системи: відповіді на питання, поставлені людською мовою.
- Розпізнавання теми: поділ тексту на частини з подальшим визначенням провідної теми для кожної з них.
- Інформаційний пошук: пошук, розпізнавання та видобування інформації.
- Добування даних: отримання семантичної інформації з тексту.

- Отримання зв’язків: визначення відносин між об'єктами у певному шматку тексту.
- Спрощення тексту: обробка інформації для спрощення структури тексту зі збереженням основної думки.
- Розв'язання лексичної багатоманітності: надання списку можливих значень конкретного багатозначного слова, серед яких можна вибрати найбільш підходяще відповідно до контексту.
- Розпізнавання абревіатур та заголовків
- Детектування окремих лінгвістичних одиниць
- Морфологічна декомпозиція: перетворення окремих термінів у зрозумілу форму.

В цьому розділі також розглянуто підходи до вирішення завдань обробки природньої мови, а також головні напрями їх реалізації. До таких входять:

- Статистичних підхід
- Лінгвістичний підхід
- Символічний підхід
- Конективістський підхід
- Метод допоміжних векторів
- Прихована марківська модель
- Умовні випадкові поля
- N-грамні моделі

У другому розділі здійснено аналіз та дослідження сучасного стану систем електронного перекладу. Розглянуто типи та різновидності електронних перекладачів, їх можливості, ознаки та особливості. Досліджені теоретичні питання побудови систем електронного перекладу. Розглянуті можливості систем штучного інтелекту при реалізації систем електронного перекладу.

Обробка природної мови (NLP) - це здатність машин розуміти та інтерпритувати людську мову. Мета NLP - заповнити прогалину між тим, як люди спілкуються (природна мова) і тим, що розуміє комп'ютер (машинна мова).

В основі систем обробки природньої мови лежить лінгвістичний аналіз, до якого входять:

- *Синтаксис* - яка частина поданого тексту є граматично правильною.
- *Семантика* - у чому сенс поданого тексту?
- *Прагматика* - яка мета тексту?

Технології NLP мають справу з різними аспектами мови, такими як:

- Фонологія - це систематична організація звуків у мові.
- Морфологія - це дослідження словотворення та їх взаємозв'язку між собою.

Загальна схема складових обробки природньої мови представлена на рис.2.1

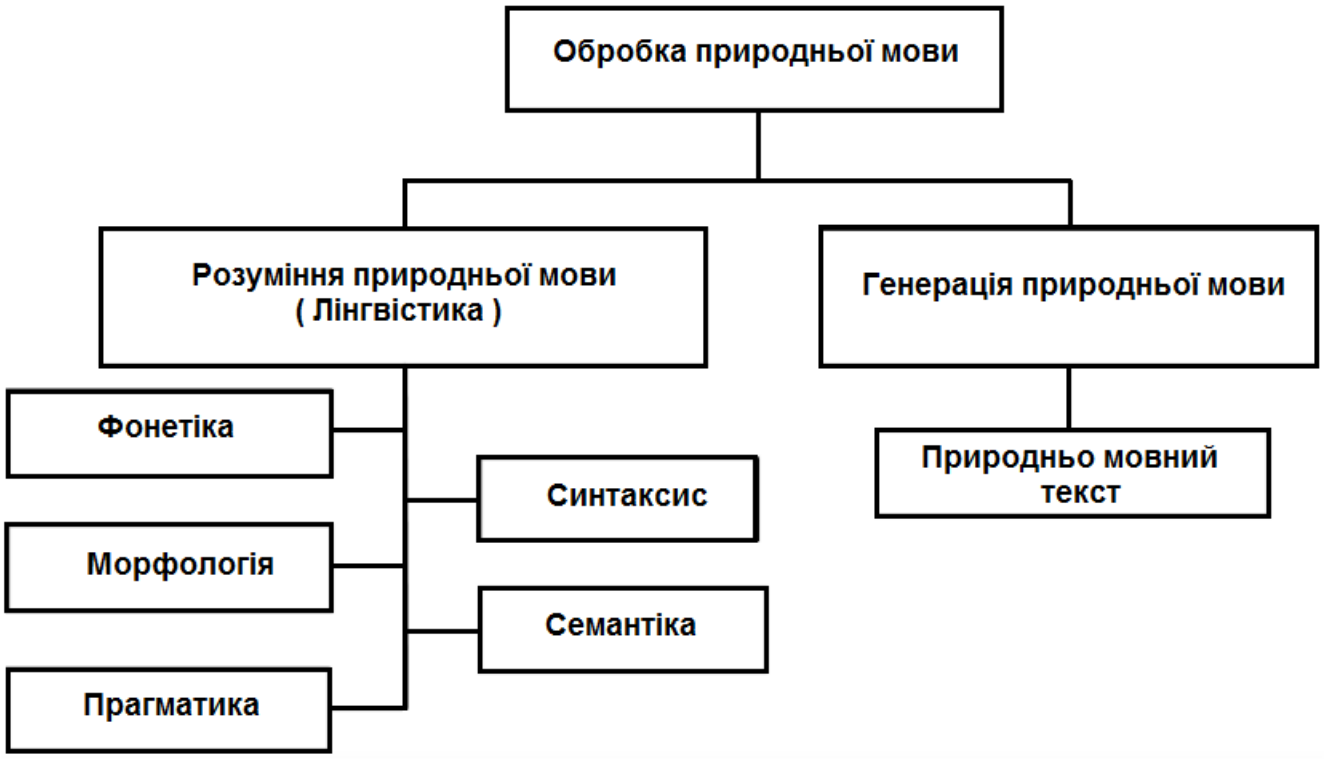


Рис.2.1 Схема складових обробки природньої мови

Інтерактивне навчання - прагматичний підхід, при якому користувач відповідає за навчання комп'ютера поетапному вивченню мови в інтерактивному навчальному середовищі.

Механізм обробки природної мови включає наступні процеси:

- Розуміння природної мови
- Генерацію природної мови

Розуміння природної мови. NLU або Natural Language Understanding намагається зрозуміти значення даного тексту. Характер та структура кожного слова в тексті повинна бути відома для NLU. Для розуміння структури NLU намагається вирішити неднозначності присутні в природній мові. Сенс кожного слова розуміється за допомогою лексикон (лексики) та набору граматичних правил. Однак деякі слова мають схоже значення (синоніми), або слова, що мають більше одного значення (полісемія).

Генерація природної мови - це процес автоматичного створення тексту зі структурованих даних у зручному для читання форматі із значущими фразами та реченнями. Проблему генерації природних мов важко вирішити. Це підмножина NLP.

Породження природної мови поділено на три запропоновані етапи:

Планування тексту - виконується впорядкування первинного вмісту в структурованих даних.

Планування речень - речення поєднуються зі структурованими даними для відображення потоку інформації.

Реалізація – отримання граматично правильних речення, що представляють текст.

Системи електронного перекладу – системи приизначені для синтезу інформацію, з використання таблиць посилань. Існують два основні типи систем перекладу - словники та перекладачі. Словники дозволяють шукати слово та знаходити його еквівалент іншою мовою. Вони не дають уявлення про структуру чи правила мови, тому, функціональні можливості їх обмежені. Перекладачі в свою чергу організовують слова навколо контексту, що допомагає краще розуміти слова та дозволяє бачити їх належне використання в фразах та реалістичному до життся тесті.

Сучасні перекладачі мають три основні конфігурації на основі того, як користувач вводить запит і як пристрій повертає результати. До них входять:

- Немовні перекладачі
- Перекладачі тексту в мову
- Перекладачі з мови в мову

Електронні перекладачі зберігають багато готових фраз в пам'яті пристрою, що дозволяє їм швидко відповідати на запити користувачів. Ці пристрої перетворюють аналогові звукові хвилі в цифровий формат, розбивають слова на фонеми, а потім порівнюючи фонеми зі словником щоб знаходять найкращий збіг. Озвучуючи інформацію вони використовують технологію синтезу мовлення або технологію TTS у два етапи. Спочатку вони аналізують слово, потім використовуючи попередньо записані відповідні збирають та відтворюють аудіофайл. Найважливішою частиною цих пристроїв роботою цих пристроїв є їх алгоритми. Все починається з бази даних паралельних текстів двома різними мовами. Далі, набір операцій визначає короткі фрази, що відповідають джерелам, і вимірює, як часто і де слова трапляються в даній фразі в обох мовах. Нарешті, програмне забезпечення використовує цю інформацію для побудови статистичних моделей, які пов'язують фрази однією мовою із фразами другої. Електронний перекладач використовує подібні обчислення, коли користувач детально аналізує категорію фраз і говорить або вводить фразу. Комп’ютер аналізує введені дані, знаходить збіг високої ймовірності та повертає результати.

Остаточна особливість перекладачів – це забезпечення повноцінний машинного перекладу, який дозволяє вводити текст і перекладати його за допомогою статистичних методів.

Сучасні системи штучного інтелекту постійно удосконалюються і розвиваються. Широке застосування знаходять моделі штучного інтелекту, побудовані за зразком живого мозку - так звані штучні нейронні мережі. Дані системи штучного інтелекту відносяться до моделей машинного навчання та використовуються, в тому числі в програмах для автоматичного перекладу текстів. Нейронні мережі таких інтелектуальних систем здатні навчатися і адаптуватися - на

прикладах перекладів текстів можна навчити нейронну мережу перекладати текст з англійської мови на німецьку.

Незважаючи на величезний прорив в області машинного перекладу і поява ефективних систем аналізу і перекладу текстів, проблема подолання мовного бар'єру все ще залишається невирішеною.

В третьому розділі розглядається когнітивна система IBM Watson. Тут описується структура обробки інформації, основні функції когнітивної системи та детально описується алгоритм її роботи. Також цей розділ описує напрям хмарних обчислень, платформу IBM Cloud, її структурні компоненти, сервіси та компоненти IBM Watson їх особливості та можливості.Тут також досліджені головні сервіси IBM Watson та інші компоненти IBM Watson. Розглянуто WATSON Developer Cloud Python SDK.

IBM Watson - хмарна платформа когнітивних обчислень, що застосовується в широкому спектрі реальних сценаріїв та завдань. Системи когнітивних обчислень моделюють можливості розподілу закономірностей та прийняття рішення людського мозку для «навчання» в ході споживання великих об'ємів даних.

Основні функції мовної когнітивної системи Watson

Система працює в такому порядку:

1. Отримання питання, Watson виконує його синтаксичний аналіз, щоб виділити основні особливості питань.

2. Система генерує ряд гіпотез, проглядаючи текст у пошукових фразах, які з кількома долями ймовірностей можуть містити необхідну відповідь. Для того, щоб вести ефективний пошук у потоках неструктурованої інформації, необхідні вдосконалені інші можливості системи їх називають когнітивними системами.

3. Система виконує глибоке порівняння мови питання та мови кожного з можливих варіантів відповідей, застосовуючи різні алгоритми логічного виводу. Це складний етап. Існують сотні алгоритмів логічного висновку, і всі вони виконують різні порівняння.

4. Кожен алгоритм логічного висновку виставляє одну або кілька оцінок, які показують, якою мірою можлива відповідь випливає із запитання.

5. Кожній отриманій оцінці потім присвоюється ваговий коефіцієнт за статистичною моделлю, яка фіксує, наскільки успішно впорався алгоритм з виявленням логічних зв'язків між двома аналогічними фразами з цієї області в "період навчання" Watson. Ця статистична модель може бути використана згодом для визначення загального рівня впевненості системи Watson в тому, що можливий варіант відповіді випливає із запитання.

6. Watson повторює процес для кожного можливого варіанту відповіді до тих пір, поки не знайде відповіді, які будуть мати більше шансів опинитися правильними, ніж інші.

Загальний зміст тексту *Watson* виводить з отриманої інформації, з додатковою бази. Когнітивні системи, їх способи збору, запам'ятовування і добування інформації схожі з тим, як аналізує інформацію людина. При цьому когнітивні системи можуть

передавати інформацію та діяти. Ось приклади поведінкових конструктів, які використовуються в цьому випадку:

- здатність створювати і перевіряти гіпотези;

- здатність розбивати на складові і будувати логічні висновки про мову;

- здатність отримувати і оцінювати корисну інформацію (таку як дати, місця розташування і характеристики).

Без цих здібностей ні комп'ютер, ні людина не зможуть визначити правильну взаємозв'язок між питаннями і відповідями. Когнітивні процеси вищого порядку можуть досягти високого рівня розуміння, орієнтуючись на основні способи поведінки. Для того щоб зрозуміти щось, ми повинні вміти розділити інформацію на більш дрібні елементи, які досить добре впорядковані на даному рівні. Фізичні процеси у людини протікають зовсім не так, як процеси в космічному масштабі або на рівні елементарних частинок. Так само і когнітивні системи призначені для роботи на рівні людини, хоча вони представляють безліч людей.

У зв'язку з цим розуміння мови починається з розуміння більш простих правил мови - не тільки формальної граматики, а й неформальних угод, які спостерігаються в повсякденному використанні.

Хмарні обчислення - це модель, яка забезпечує зручний мережевий доступ до спільного пулу конфігурованих обчислювальних ресурсів, які можна швидко забезпечити та випустити з мінімальними зусиллями з боку керівництва, або взаємодії з постачальником послуг

Хмарні обчислення надають наступні переваги для розробників:

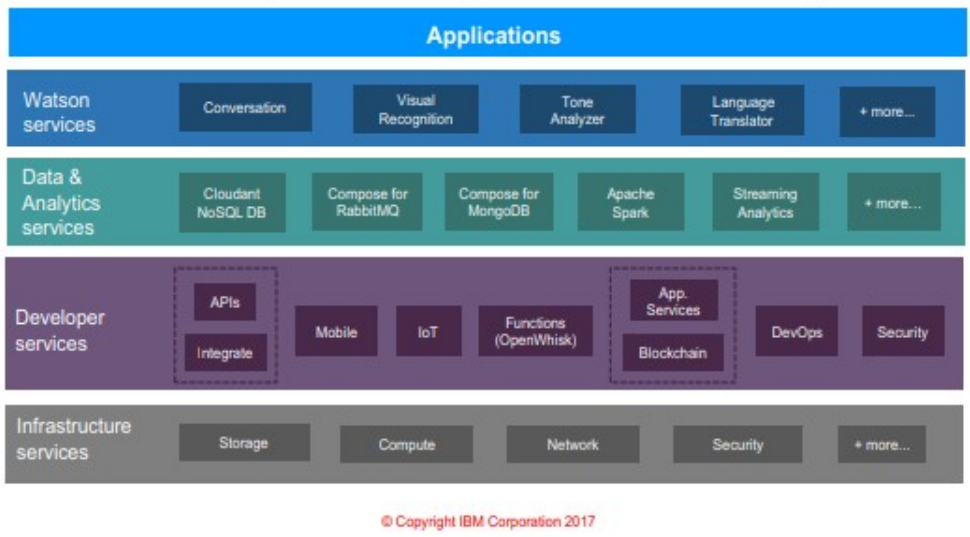
- Легко доступний інструментарій та виробниче середовище.
- Безкоштовні пробні версії більшості продуктів.
- Попередньо побудовані шаблони сприяють швидко розпочати роботу.
- Простіше зрозуміти життєвий цикл програмного забезпечення (ПЗ).
- Навколишнє середовище встановлюється за хвилини, а не за дні.

IBM Cloud - це відкрита платформа хмарних обчислень, яка поєднує платформу як сервіс (*PaaS*) з інфраструктурою як послугою (*IaaS*), і включає каталог різноманітних хмарних сервісів, які можуть використовуватися для швидкого створення та розгортання бізнес-додатків або інфраструктури.

Структурними компонентами платформи *IBM Cloud* є:

- Надійна консоль - інтерфейс для створення, перегляду та управління хмарними ресурсами.
- Керування доступом користувачів, як для сервісів платформи, так і послідовно контролюють доступ до ресурсів у *IBM Cloud*.
- Каталог продуктів, що підтримуються *IBM Cloud*
- Механізм пошуку та позначення для фільтрації та ідентифікації ресурсів.
- Сховища даних.
- Безпечна система управління рахунками та виставленням рахунків.

IBM Cloud надає широкий спектр готових послуг який можна використовувати при складанні програми. Сервіси Watson дозволяють додати потужність штучного інтелекту до любого програмного додатку API для обробки мови, зору та природної мови.



Сервіси IBM Cloud

У четвертому розділі наводиться опис розробки програмного забезпечення системи перекладу, алгоритмтм її роботи, структура та реалізація мобільного

додатку. Тут також описане створення словника перекладача, описані конфігурація, особливості налаштування та роботи використаних сервісів IBM Watson та платформи IBM Cloud.

У спеціальній частині магістерської кваліфікаційної роботи з «Охорони праці та безпеки життєдіяльності» » було проведено аналіз умов праці в розглянутому робочому приміщенні який показав, що умови праці з ПЕОМ відповідають вимогам. Приміщення, в якому розташовано робоче місце має достатні площу та об’єм для роботи однієї людини. Завдяки використанню сучасного обладнання та підбору оптимальної продуктивності комп’ютера відповідно до роботи, що виконується, рівень шуму комп’ютера не перевищує нормативні значення. Аналіз параметрів мікроклімату показав, що вони не в повній мірі відповідають вимогам нормативних документів. В зимовий період вологість повітря знаходиться на межі допустимих значень, а в теплу пору року температура повітря на робочому місці може перевищувати норму. Для приведення мікроклімату до відповідності нормам необхідне застосування зволожувача повітря у зимовий період, та кондиціонера у теплу пору року. Ергономіка робочого місця і режим зорової роботи задовольняють вимогам і сприяють зниженню втоми. Для збереження здоров'я робітника, запобігання професійним захворювання і підтримки працездатності необхідно суворо дотримуватись вимог до режимів праці і відпочинку при роботі з ВДТ ЕОМ і ПЕОМ. Також було приведено основні рекомендації для роботодавців та працівників, щодо заходів під час епідемії COVID-19.

У методичній частині було розлянуто метод аналізу ієрархій з використанням тверджень декількох експертів, а саме алгоритм реалізації цього метода для двох випадків: при проведенні опитування серед рівноцінних експертів і при проведенні опитування серед експертів з різними вагами.

ЗАГАЛЬНІ ВИСНОВКИ

У даній магістерські кваліфікаційній роботі досліджена інтелектуальна система on-line перекладу на базі хмарних сервісів.

У першому розділі наведено огляд проблем обробки та структуризації людського мовлення. Розглянуто обробку природньої мови та єтапи її історичного розвитку.

У другому розділі здійснено аналіз та дослідження сучасного стану систем електронного перекладу. Розглянуто типи та різновидності електронних перекладачів, їх можливості, ознаки та особливості Досліджені теоретичні питання побудови систем електронного перекладу. Розглянуті можливості систем штучного інтелекту при реалізації систем електронного перекладу.

В третьому розділі розглядається когнітивна система IBM Watson. Тут описується структура обробки інформації, основні функції когнітивної системи та детально описується алгоритм її роботи. Також цей розділ описує напрям хмарних обчислень, платформу IBM Cloud, її структурні компоненти, сервіси та компоненти IBM Watson їх особливості та можливості.

У четвертому розділі описаний процес розробки програмного забезпечення системи перекладу.

У методичній частині було розлянуто метод аналізу ієрархій з використанням тверджень декількох експертів,

У спеціальній частині магістерської кваліфікаційної роботи з «Охорони праці та безпеки життєдіяльності» » було проведено аналіз умов праці в розглянутому робочому приміщенні.

АНОТАЦІЯ

Мельничук Іван Олегович. Інтелектуальна система on-line перекладу на базі хмарних сервісів.

Дипломна робота на здобуття освітньої кваліфікації «Магістр комп'ютерних наук». – Чорноморський національний університет імені Петра Могили, Миколаїв, 2021.

Дана робота присвячена дослідженню методів, алгоритмів обробки природньої мови на основі когнітивного хмарного сервісу IBM Watson. У рамках магістерської роботи був реалізований електронний технічний перекладач на основі мовних сервісів IBM Watson.

Об'єкт дослідження – процеси обробки природньої мови в хмарних сервісах.
Мета роботи – методи та сервіси обробки природньої мови в хмарних сервісах.

Практичне значення результатів дослідження полягає у можливості їх використання для створення інтелектуальних мовних систем різного призначення.

Дипломна робота складається з фахового розділу, методичної і спеціальної частини з охорони праці.

Фахова частина записка магістерської роботи складається зі вступу, чотирьох розділів, висновків та двох додатків. У вступі визначається актуальність теми та проводиться короткий огляд поставленої задачі.

У першому розділі розглядається задачі і методи обробки природньої мови; аналізуються підходи до проблеми побудови інтелектуальних мовних додатків. Та побудови систем обробки природньої мови. У другому розділі проводиться аналіз та дослідження систем електронного перекладу. Розглянуті питання класифікації та побудови систем електронного перекладу. А також досліджені питання використання штучного інтелекту при розробці мовних додатків. У третьому розділі докладніше описується когнітивна система IBM Watson, а реалізація хмарних сервісів та детально описуються мовні сервіси IBM Watson. В четвертому розділі описується процес проектування програмного забезпечення технічного перекладача; наведений опис алгоритму роботи перекладача. Та процес взаємодії мовних сервісів та перекладача. Також описано мобільний додаток і його реалізація для зв'язку з хмарною платформою Watson.

У висновках проводиться аналіз проведеної роботи та отриманих результатів.

В спеціальній частині з охорони праці розглянуто питання безпеки на робочому місті. Методична частина містить методичні матеріали до виконання практичної роботи з курсу Методи і системи штучного інтелекту.

В цілому, магістерська наукова робота без додатків містить 92 сторінки, 17 рисунків.

ABSTRACT

Melnychuk Ivan Olehovych. Intelligent on-line translation system based on cloud services.

Thesis for the educational qualification "Master of Computer Science". - Petro Mohyla Black Sea National University, Mykolaiv, 2021.

This work is devoted to the study of methods, algorithms for natural language processing based on IBM Watson cognitive cloud service. As part of the master's thesis, an electronic technical translator based on IBM Watson language services was implemented.

The object of study - the processes of natural language processing in cloud services. The purpose of the work - methods and services of natural language processing in cloud services.

The practical significance of the research results lies in the possibility of their use to create intelligent language systems for various purposes.

Thesis consists of a professional section, methodical and special part on labor protection.

The professional part of the master's thesis note consists of an introduction, four sections, conclusions and two appendices. The introduction determines the relevance of the topic and provides a brief overview of the task.

The first section discusses the tasks and methods of natural language processing; approaches to the problem of building intelligent language applications are analyzed. And building natural language processing systems. The second section analyzes and studies electronic translation systems. The issues of classification and construction of electronic translation systems are considered. And also questions of use of artificial intelligence at development of language applications are investigated. The third section describes the IBM Watson cognitive system in more detail, and the implementation of cloud services, and describes the IBM Watson language services in detail. The fourth section describes the process of designing technical translator software; the description of the algorithm of work of the translator is given. And the process of interaction between language services and translator. The mobile application and its implementation for communication with the Watson cloud platform are also described.

The conclusions analyze the work done and the results obtained.

In the special part on labor protection the issues of safety at the workplace are considered. The methodical part contains methodical materials for performance of practical work from a course Methods and systems of artificial intelligence.

In general, the master's thesis without appendices contains 92 pages, 17 figures.