Хмельницький національний університет
Факультет інформаційних технологій
Кафедра комп'ютерних наук

КВАЛІФІКАЦІЙНА РОБОТА
МАГІСТРА

на тему *Метод класифікації текстових документів засобами машинного навчання*

Галузь знань *12 – Інформаційні технології*
Шифр і назва галузі знань

Спеціальність *122 – Комп'ютерні науки*
Шифр і назва спеціальності

Виконав:	студентка *2* курсу, групи *КНм-20-2*	Підпис	*Т.К. Скрипник* Ініціали, прізвище
Керівник:	*д.т.н., професор кафедри КН*	Підпис	*О.В. Бармак* Ініціали, прізвище
Нормоконтроль:	*к.т.н., доцент кафедри КН*	Підпис	*Р.О. Багрій* Ініціали, прізвище

До захисту допускаю:
Зав. кафедри КН, д.т.н., професор Підпис *О.В. Бармак* Ініціали, прізвище

23 листопада 2021 р.

Хмельницький 2021

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

Факультет *інфомаціних технологій*
Кафедра *комп'ютерних наук*
Освітній ступінь *магістр*
Галузь знань *12 – Інформаційні технології*
Спеціальність *122 – Комп'ютерні науки*

ЗАТВЕРДЖУЮ
Завідувач кафедри *комп'ютерних наук*

_____
(підпис)
д.т.н., професор О.В. Бармак
« 01 » вересня 2021 року

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ МАГІСТРА

1. Тема кваліфікаційної роботи магістра: *«Метод класифікації текстових документів засобами машинного навчання»*

2. Завдання видано студентці *Скрипник Тетяні Казимирівні*
(прізвище, ім'я, по батькові)

3. Керівник роботи *д.т.н., професор Бармак Олександр Володимирович*
(прізвище, ім'я, по батькові)

4. Затверджені наказом університету від « 25 » серпня 2021 р. № 102

5. Зміст пояснювальної записки (перелік задач) та вихідні дані:

Мета роботи – розробка методу класифікації текстових документів засобами машинного навчання.

Об'єктом дослідження є процеси отримування інформації з використанням методів машинного навчання.

Предметом дослідження є моделі та методи обробки текстових даних технічної документації та перекладених документів різними мовами.

Реферат

Кваліфікаційна робота магістра присвячена розробці методу класифікації текстових документів засобами машинного навчання.

Актуальність теми. В магістерській роботі розроблено та набув практичної реалізації метод класифікації технічної документації на основі перекладу речень.

Можливість подолання мовних бар'єрів є предметом інтересу з початку XVII століття. З тих пір були розроблені пристрої та ідеї для полегшення розуміння між людьми, що говорять на різних мовах, такі як механічні словники або універсальні мови. У зв'язку з високою інтернаціоналізацією компаній і загальною глобалізацією, можливість автоматичного перекладу текстів з однієї мови на іншу без допомоги людини є темою, якою займаються вже близько 60 років, а в останнє десятиліття вона набула ще більшого інтересу. Щоб успішно працювати на міжнародному ринку, компанії повинні надавати добре перекладену документацію на свою продукцію. Оскільки складні продукти часто мають більше однієї групи користувачів, наприклад, адміністраторів, користувачів і розробників, кількість різних документів для одного продукту може бути дуже великою. Особливо для компаній, орієнтованих на експорт, важко знайти професійних перекладачів з відповідною технічною освітою для створення правильно перекладеної технічної документації за розумною ціною. Тому рішення машинного перекладу викликають все більший інтерес у фірм. Становить великий інтерес забезпечення автоматизованого високоякісного перекладу текстів для забезпечення рівного доступу до інформації незалежно від мови її джерела. Зокрема, точні переклади технічної документації є пріоритетними для компаній, оскільки вони необхідні для безперебійного робочого процесу, задоволення потреб клієнтів і описують управління, функціональність, а також функції безпеки продукції. У цій області непорозуміння можуть бути дуже серйозними. Крім того, співпраця між компаніями може постраждати через непорозуміння, викликаних поганим перекладом.

Оскільки обробка природної мови є дуже складним завданням, результат, отриманий за допомогою програмного забезпечення для машинного перекладу, все

ще потребує схвалення людиною для забезпечення необхідної якості. Таким чином, просте використання програмного забезпечення для перекладу ділової документації переносить проблему з моменту створення документа на етап оцінки і коригування, але не вирішує її. Отже, оцінка перекладеної технічної документації є важливим кроком для компаній, що дозволяє скоротити час і витрати, а також створити ефективний спосіб перекладу важливих документів. Крім того, це забезпечує певний рівень якості. Складність при оцінці якості перекладу пояснюється суб'єктивним характером і різними аспектами, пов'язаними з терміном «якість», такими як граматична правильність тощо.

Доступ до комп'ютеризованих систем, які виконують правильний переклад будь-якого речення, все ще залишається далекою мрією, особливо через проблеми передачі системі сенсу. У зв'язку з цією проблемою важливо мати можливість оцінювати якість перекладу - в іншому випадку неможливо переконатися, що документ був переведений правильно. Особливий інтерес представляє технічна документація, оскільки її велика кількість в кожній компанії, що продає продукцію, що ще більше підвищує мотивацію до автоматичного перекладу такого роду документів. В даний час компанії вирішують проблему перекладу технічної документації, передаючи цю задачу зовнішнім перекладачам. Оскільки особа, що запрошує такий переклад, не обов'язково володіє мовою перекладу, важливо переконатися, що робота була виконана правильно і професійно.

Метою дослідження є розробка методу реалізації процесу машинного навчання, який здатний класифікувати текстові дані, та визначити чи були документи перекладені професійними перекладачами або комп'ютерними системами трансляції текстової інформації.

Для досягнення зазначеної мети поставлені наступні **задачі**:

– визначити мету процесу і зібрати попередні необхідні знання про прикладну область;

– вибрати відповідний набір даних для отримання знань;

– попередня обробка даних: видалення шумів або шкідливих записів даних

та прийняття рішення про певні налаштування, наприклад, про те, як обробляти відсутні значення атрибутів в наборі даних;

– привести дані в прийнятний формат, наприклад, видалити непотрібні змінні або параметри з точки зору мети завдання;

– прийняти рішення про підхід до отримання даних для певної мети процесу отримання знань;

– після прийняття рішення про загальний підхід до аналізу даних наступним кроком є вибір алгоритму аналізу даних: важливо відзначити, що цей вибір часто залежить від уподобань кінцевого користувача, наприклад, перевага віддається зрозумілому формату або максимальної якості прогнозування;

– основний етап отримання даних, який полягає в застосуванні алгоритму до попередньо обробленого набору даних та пошуку цінних знань у даних;

– інтерпретувати знайдені алгоритмом закономірності і, можливо, повернутися до одного з попередніх етапів, щоб скорегувати настройку процесу отримання знань;

– використання інтерпретованих результатів для подальших дій, наприклад, для подальшого дослідження або застосування систем до реального сценарію.

Об'єктом дослідження є процеси отримування інформації з використанням методів машинного навчання.

Предметом дослідження є моделі та методи обробки текстових даних технічної документації та перекладених документів різними мовами.

Наукова новизна одержаних результатів. В результаті проведеної роботи були отримані такі результати:

– набула подальшого розвитку система оцінки якості класифікації текстової інформації, яка специфікується областю дослідження та є інтегральним показником якості класифікації;

– запропоновано інноваційний метод визначення якості перекладу текстової технічної документації на основі класифікації перекладу виконаного перекладачами та автоматизованими системами перекладу;

– запропоновано метод визначення якості перекладу із застосуванням методів розмічених та нерозмічених даних.

В умовах все більш мережевого світу наявність високоякісних перекладів має вирішальне значення для успіху в умовах зростаючої міжнародної конкуренції. Великі міжнародні компанії, а також компанії середнього розміру повинні надавати своїм клієнтам добре перекладену технічну документацію високої якості не тільки для того, щоб бути успішними на ринку, але і для того, щоб відповідати правовим нормам і уникнути судових позовів.

Тому дана робота присвячена оцінці якості перекладу, зокрема, технічної документації, і відповідає на два основних питання:

– як можна оцінити якість перекладу технічної документації, якщо є оригінал документа;

– як можна оцінити якість перекладу технічної документації, якщо оригінал документа недоступний.

Відповіді на ці питання даються за допомогою сучасних алгоритмів машинного навчання і метрик оцінки перекладу в контексті процесу виявлення знань. Оцінка проводиться на рівні пропозицій і об'єднується на рівні документів шляхом бінарної класифікації пропозицій на автоматичний і професійний переклад. Дослідження засноване на базі даних. На основі розроблених систем класифікації за пропозиціями, документи класифікуються за допомогою рекомбінації пов'язаних пропозицій, а також вводиться основа для оцінки якості документів.

Достовірність результатів забезпечується використанням сучасних методів та засобів машинного навчання і штучного інтелекту та напрацювань в межах області дослідження.

Доступ до комп'ютеризованих систем, які виконують правильний переклад будь-якого речення, все ще залишається далекою мрією, особливо через проблеми передачі системі сенсу. У зв'язку з цією проблемою важливо мати можливість оцінювати якість перекладу - в іншому випадку неможливо переконатися, що документ був перекладений правильно. Особливий інтерес представляє технічна

документація, оскільки її велика кількість в кожній компанії, що продає продукцію, що ще більше підвищує мотивацію до автоматичного перекладу такого роду документів. В даний час компанії вирішують проблему перекладу технічної документації, передаючи цю задачу зовнішнім перекладачам. Оскільки особа, що запрошує такий переклад, не обов'язково володіє мовою перекладу, важливо переконатися, що робота була виконана правильно і професійно.

У роботі запропонований інноваційний метод визначення якості перекладу технічної документації з використанням методів машинного навчання. Метод базується на моделях машинного навчання, розрахунку метрик якості класифікації перекладеної документації.

Практична значимість дослідження полягає в тому, що отримані практичні результати досліджень можуть бути застосовні для визначення якості перекладу технічної документації.

Розроблена система оцінки для ранжирування пропозицій і документів на основі їх якості незалежно від типу перекладу. Запропонована модель складається з методів, які використовують дві оптимізовані моделі машинного навчання для класифікації речень і додатковий незалежний від посилань інструмент перевірки граматики і орфографії для створення виваженої кількості помилок для кожного речення. Для оцінки якості документа класи якості відповідних речень усереднюються з додатковою вагою для зменшення кількості помилок.

У даній роботі представлена система класифікації технічних документів з використанням методів машинного навчання і підхід до оцінки якості документів. У продовження цієї теми можна поліпшити якість класифікації на рівні документа, об'єднавши речення на рівні документа на основі обчисленої достовірності результуючих моделей класифікації. Такий підхід призведе до більш тонкої класифікації документів, оскільки буде враховуватися впевненість алгоритму в класифікації на основі речень.

Апробація кваліфікаційної роботи.

Основні положення і результати роботи опубліковані:

Аналітична система визначення якості перекладу текстової інформації методами машинного навчання / Скрипник Т.К., Манзюк Е.А. // Збірник наукових праць за матеріалами Міжнародної конференції «IX Українсько-Польські наукові діалоги», Хмельницький, Україна 20-23 жовтня 2021 р., – С.151-153.

Структура та обсяг роботи. Кваліфікаційна робота магістра складається з завдання, реферату, змісту, вступу, 4 розділів, висновків, переліку посилань із 28 найменувань та додатків. Загальний обсяг кваліфікаційної роботи магістра становить 123 сторінки, з них 86 сторінок основного тексту та 37 сторінки додатків. В роботі наведено 14 рисунків та 13 таблиць.

Ключові слова: машине навчання, класифікація, метрики оцінки якості.

Зміст

Вступ

Кваліфікаційна робота магістра присвячена розробці методу класифікації текстових документів засобами машинного навчання.

Актуальність теми. В магістерській роботі розроблено та набув практичної реалізації метод класифікації технічної документації на основі перекладу речень.

Можливість подолання мовних бар'єрів є предметом інтересу з початку XVII століття. З тих пір були розроблені пристрої та ідеї для полегшення розуміння між людьми, що говорять на різних мовах, такі як механічні словники або універсальні мови. У зв'язку з високою інтернаціоналізацією компаній і загальною глобалізацією, можливість автоматичного перекладу текстів з однієї мови на іншу без допомоги людини є темою, якою займаються вже близько 60 років, а в останнє десятиліття вона набула ще більшого інтересу. Щоб успішно працювати на міжнародному ринку, компанії повинні надавати добре перекладену документацію на свою продукцію. Оскільки складні продукти часто мають більше однієї групи користувачів, наприклад, адміністраторів, користувачів і розробників, кількість різних документів для одного продукту може бути дуже великою. Особливо для компаній, орієнтованих на експорт, важко знайти професійних перекладачів з відповідною технічною освітою для створення правильно перекладеної технічної документації за розумною ціною. Тому рішення машинного перекладу викликають все більший інтерес у фірм. Становить великий інтерес забезпечення автоматизованого високоякісного перекладу текстів для забезпечення рівного доступу до інформації незалежно від мови її джерела. Зокрема, точні переклади технічної документації є пріоритетними для компаній, оскільки вони необхідні для безперебійного робочого процесу, задоволення потреб клієнтів і описують управління, функціональність, а також функції безпеки продукції. У цій області непорозуміння можуть бути дуже серйозними. Крім того, співпраця між компаніями може постраждати через непорозуміння, викликаних поганим перекладом.

Оскільки обробка природної мови є дуже складним завданням, результат, отриманий за допомогою програмного забезпечення для машинного перекладу, все ще потребує схвалення людиною для забезпечення необхідної якості. Таким чином, просте використання програмного забезпечення для перекладу ділової документації переносить проблему з моменту створення документа на етап оцінки і коригування, але не вирішує її. Отже, оцінка перекладеної технічної документації є важливим кроком для компаній, що дозволяє скоротити час і витрати, а також створити ефективний спосіб перекладу важливих документів. Крім того, це забезпечує певний рівень якості. Складність при оцінці якості перекладу пояснюється суб'єктивним характером і різними аспектами, пов'язаними з терміном «якість», такими як граматична правильність тощо.

Доступ до комп'ютеризованих систем, які виконують правильний переклад будь-якого речення, все ще залишається далекою мрією, особливо через проблеми передачі системі сенсу. У зв'язку з цією проблемою важливо мати можливість оцінювати якість перекладу - в іншому випадку неможливо переконатися, що документ був перекладений правильно. Особливий інтерес представляє технічна документація, оскільки її велика кількість в кожній компанії, що продає продукцію, що ще більше підвищує мотивацію до автоматичного перекладу такого роду документів. В даний час компанії вирішують проблему перекладу технічної документації, передаючи цю задачу зовнішнім перекладачам. Оскільки особа, що запрошує такий переклад, не обов'язково володіє мовою перекладу, важливо переконатися, що робота була виконана правильно і професійно.

Метою дослідження є розробка методу реалізації процесу машинного навчання, який здатний класифікувати текстові дані, та визначити чи були документи перекладені професійними перекладачами або комп'ютерними системами трансляції текстової інформації.

Для досягнення зазначеної мети поставлені наступні **задачі**:

– визначити мету процесу і зібрати попередні необхідні знання про прикладну область;

– вибрати відповідний набір даних для отримання знань;

– попередня обробка даних: видалення шумів або шкідливих записів даних та прийняття рішення про певні налаштування, наприклад, про те, як обробляти відсутні значення атрибутів в наборі даних;

– привести дані в прийнятний формат, наприклад, видалити непотрібні змінні або параметри з точки зору мети завдання;

– прийняти рішення про підхід до отримання даних для певної мети процесу отримання знань;

– після прийняття рішення про загальний підхід до аналізу даних наступним кроком є вибір алгоритму аналізу даних: важливо відзначити, що цей вибір часто залежить від уподобань кінцевого користувача, наприклад, перевага віддається зрозумілому формату або максимальної якості прогнозування;

– основний етап отримання даних, який полягає в застосуванні алгоритму до попередньо обробленого набору даних та пошуку цінних знань у даних;

– інтерпретувати знайдені алгоритмом закономірності і, можливо, повернутися до одного з попередніх етапів, щоб скорегувати настройку процесу отримання знань;

– використання інтерпретованих результатів для подальших дій, наприклад, для подальшого дослідження або застосування систем до реального сценарію.

Об'єктом дослідження є процеси отримування інформації з використанням методів машинного навчання.

Предметом дослідження є моделі та методи обробки текстових даних технічної документації та перекладених документів різними мовами.

Наукова новизна одержаних результатів. В результаті проведеної роботи були отримані такі результати:

– набула подальшого розвитку система оцінки якості класифікації текстової інформації, яка специфікується областю дослідження та є інтегральним показником якості класифікації;

– запропоновано інноваційний метод визначення якості перекладу текстової технічної документації на основі класифікації перекладу виконаного перекладачами та автоматизованими системами перекладу;

– запропоновано метод визначення якості перекладу із застосуванням методів розмічених та нерозмічених даних.

В умовах все більш мережевого світу наявність високоякісних перекладів має вирішальне значення для успіху в умовах зростаючої міжнародної конкуренції. Великі міжнародні компанії, а також компанії середнього розміру повинні надавати своїм клієнтам добре перекладену технічну документацію високої якості не тільки для того, щоб бути успішними на ринку, але і для того, щоб відповідати правовим нормам і уникнути судових позовів.

Тому дана робота присвячена оцінці якості перекладу, зокрема, технічної документації, і відповідає на два основних питання:

– як можна оцінити якість перекладу технічної документації, якщо є оригінал документа;

– як можна оцінити якість перекладу технічної документації, якщо оригінал документа недоступний.

Відповіді на ці питання даються за допомогою сучасних алгоритмів машинного навчання і метрик оцінки перекладу в контексті процесу виявлення знань. Оцінка проводиться на рівні речень і об'єднується на рівні документів шляхом бінарної класифікації речень на автоматичний і професійний переклад. Дослідження засноване на базі даних. На основі розроблених систем класифікації за реченнями, документи класифікуються за допомогою рекомбінації пов'язаних пропозицій, а також вводиться основа для оцінки якості документів.

Достовірність результатів забезпечується використанням сучасних методів та засобів машинного навчання і штучного інтелекту та напрацювань в межах області дослідження.

Доступ до комп'ютеризованих систем, які виконують правильний переклад будь-якого пречення, все ще залишається далекою мрією, особливо через проблеми

передачі системі сенсу. У зв'язку з цією проблемою важливо мати можливість оцінювати якість перекладу - в іншому випадку неможливо переконатися, що документ був перекладений правильно. Особливий інтерес представляє технічна документація, оскільки її велика кількість в кожній компанії, що продає продукцію, що ще більше підвищує мотивацію до автоматичного перекладу такого роду документів. В даний час компанії вирішують проблему перекладу технічної документації, передаючи цю задачу зовнішнім перекладачам. Оскільки особа, що запрошує такий переклад, не обов'язково володіє мовою перекладу, важливо переконатися, що робота була виконана правильно і професійно.

У роботі запропонований інноваційний метод визначення якості перекладу технічної документації з використанням методів машинного навчання. Метод базується на моделях машинного навчання, розрахунку метрик якості класифікації перекладеної документації.

Практична значимість дослідження полягає в тому, що отримані практичні результати досліджень можуть бути застосовні для визначення якості перекладу технічної документації.

Розроблена система оцінки для ранжирування речень і документів на основі їх якості незалежно від типу перекладу. Запропонована модель складається з методів, які використовують дві оптимізовані моделі машинного навчання для класифікації пропозицій і додатковий незалежний від посилань інструмент перевірки граматики і орфографії для створення виваженої кількості помилок для кожного речення. Для оцінки якості документа класи якості відповідних речень усереднюються з додатковою вагою для зменшення кількості помилок.

У даній роботі представлена система класифікації технічних документів з використанням методів машинного навчання і підхід до оцінки якості документів. У продовження цієї теми можна поліпшити якість класифікації на рівні документа, об'єднавши речення на рівні документа на основі обчисленої достовірності результуючих моделей класифікації. Такий підхід призведе до більш тонкої

класифікації документів, оскільки буде враховуватися впевненість алгоритму в класифікації на основі речень.

Апробація кваліфікаційної роботи.

Основні положення і результати роботи опубліковані:

Аналітична система визначення якості перекладу текстової інформації методами машинного навчання / Скрипник Т.К., Манзюк Е.А. // Збірник наукових праць за матеріалами Міжнародної конференції «IX Українсько-Польські наукові діалоги», Хмельницький, Україна 20-23 жовтня 2021 р., – С. 151- 153.

Структура та обсяг роботи. Кваліфікаційна робота магістра складається з завдання, реферату, змісту, вступу, 4 розділів, висновків, переліку посилань із 28 найменувань та додатків. Загальний обсяг кваліфікаційної роботи магістра становить 123 сторінки, з них 86 сторінок основного тексту та 37 сторінки додатків. В роботі наведено 14 рисунків та 13 таблиць.

Розділ 1
Системи машинного навчання при класифікації текстової інформації

1.1 Опис предметної області

Як згадувалося вище, ідея автоматичного рейтингування машинних перекладів не нова. Одним з основних методів, що обговорюються в літературі, є "переклад туди і назад", який працює шляхом перекладу фрагмента тексту на іноземну мову і назад. Після цього оригінальний текст і новостворений текст порівнюються. На цій основі алгоритм "двомовної оцінки дублера" (bilingual evaluation understudy АДО) був розроблений в 2002 році. АДО визначає якість документа як сильну кореляцію між машинним перекладом і роботою професійних перекладачів-людей. В його основі лежить ідея про те, що крім точності перекладу слів важлива довжина перекладеного тексту. За даними [1], переклади людиною мають тенденцію бути більш якісними і короткими, ніж комп'ютерні. Ця ідея отримала подальший розвиток в США в вигляді алгоритму NIST для оцінки машинного перекладу, розробленого Національним інститутом стандартів і технологій. Цей алгоритм зважує чи збігаються слова відповідно до їх частоти у відповідному еталонному перекладі. Другою метрикою АДО є метрика для оцінки перекладу з явним упорядкуванням, названа Meteor, яка була розроблена Лави та ін. у 2005 році [2]. Основна відмінність Meteor полягає в здатності визначати синоніми слів, що призводить до потенційно менш помилкового перекладу. Крім того [3] пропонують використовувати допоміжні SVM для класифікації машинних перекладів на рівні речень. Розширюючи оцінку на рівні речень, були проведені додаткові дослідження по заміні використання референсних перекладів, які часто вимагають великих ресурсів для оцінки систем машинного перекладу. Пропонують використовувати ймовірності лексикону [4], також пропонують використовувати псевдо посилання [5], замінюючи часто використовувані людські еталонні переклади декількома системами автоматичного перекладу як еталони і комбінуючи

обчислені бали порівнювати з класифікатором машинного навчання для оцінки якості речень. Успішно використовують регресійне навчання в поєднанні з псевдореференсами [6, 7]. Як показано вище, по темі оцінки машинного перекладу було проведено багато досліджень.

Однак питання про зосередження уваги на конкретній області документів з метою отримання неявних додаткових знань за допомогою методів машинного навчання розглядається недостатньо, так само як і порівняння різних підходів машинного навчання для класифікації того, чи були документи переведені професійно або автоматично. Дана робота покликана відповісти на ці питання.

1.2 Інтелектуальний аналіз даних в області машинного перекладу

Проаналізуємо теоретичні основи машиного перекладу шляхом отримання відповідей на наступні питання:

– що є процесом виявлення знань про дані в базах даних;

– які переваги інтелектуального аналізу даних і які алгоритми є перспективними в області оцінки машинного перекладу;

– як працює машинний переклад і які існують підходи;

– які метрики зазвичай використовуються для вимірювання якості машинного перекладу;

– що характеризує технічну документацію.

Щоб відповісти на ці питання, розглянемо процес виявлення знань в базах даних, потім дамо визначення отримання даних, а також детально розглянемо машинне навчання в цілому і відповідні алгоритми. Також зробимо огляд сучасного стану справ в області машинного перекладу і оцінки його якості. На завершення, дамо характеристику технічної документації і опис особливостей такого роду документів.

1.3 Виявлення знань в базах даних

Процес виявлення знань в базах даних (ЗБД) описує загальний етап виявлення корисних знань з даних. Хоча існує безліч визначень процесу ЗБД, більшість з них сходяться до основних складових. У [8-10] визначають ЗБД як інтерактивний і ітеративний процес та виділяють дев'ять основних етапів.

1. Визначити мету процесу і зібрати попередні необхідні знання про прикладну область.

2. Вибрати відповідний набір даних для вилучення знань.

3. Попередня обробка даних. Сюди входить видалення шумів або шкідливих записів даних та прийняття рішення про певні налаштування, наприклад, про те, як обробляти відсутні значення атрибутів в наборі даних.

4. Привести дані в більш-менш прийнятний формат, наприклад, видалити непотрібні змінні або параметри з точки зору мети завдання.

5. Прийняти рішення про підхід до отримання даних для певної мети процесу ЗБД.

6. Після прийняття рішення про загальний підхід до аналізу даних наступним кроком є вибір алгоритму аналізу даних. Важливо відзначити, що цей вибір часто залежить від уподобань кінцевого користувача, наприклад, перевага віддається зрозумілому формату або максимальної якості прогнозування.

7. Основний етап отримання даних який полягає в застосуванні алгоритму до попередньо обробленого набору даних. Потім алгоритм шукає цінні знання в даних.

8. Інтерпретувати знайдені алгоритмом закономірності і, можливо, повернутися до одного з попередніх етапів, щоб скорегувати настройку процесу ЗБД.

9. Останнім етапом виявлення знань в базах даних є використання інтерпретованих результатів для подальших дій, наприклад, для подальшого дослідження або застосування систем до реального сценарію.

Як згадувалося в кроці 8, процес ЗБД може складатися з безлічі ітерацій і циклів. Наприклад, після інтерпретації результатів алгоритму можна зробити висновок, що обраний алгоритм був неправильним, і повернутися до кроку 5 або після приведення даних до репрезентативному формату на кроці 4 зрозуміти, що попередня обробка була виконана неправильно, і повернутися до кроку 3.

1.4 Отримання даних

Отримання даних часто використовується як синонім ЗБД. Насправді, отримання даних являє собою частину ЗБД-процесу, в якому відповідний підхід і алгоритм вибирається і застосовується до набору даних. Тому він є центральною частиною процесу виявлення знань в базах даних. Отримання даних полягає в добуванні даних в будь-якій формі і застосуванні до них алгоритмів аналізу з метою виявлення закономірностей або моделей в наборі даних і використання цих структур для класифікації даних за різними класами (мітками). Включає в себе ряд областей досліджень, таких як система баз даних, статистика і розпізнавання образів. Завдання отримання даних розрізняються залежно від знань, які алгоритм отримує про існуючі класи в наборі даних.

1. Навчання під наглядом включає в себе кожну задачу, в якій алгоритм має доступ до вхідних даних і вихідні значення. Рівні введення визначаються як зовнішня інформація, яку дозволено використовувати алгоритму, наприклад, значення атрибутів і метадані, а вихідні значення - це конкретні мітки атрибутів класу. Це означає, що структура даних вже відома, і мета цих програм - віднести нові дані до правильних класів.

2. На відміну від контрольованого навчання, неконтрольоване навчання включає всі завдання, які не мають доступу до вихідних значень і тому намагаються знайти структури в даних, створюючи класи самостійно.

Оскільки завдання представляє собою задачу бінарної класифікації, що стосується двох класів – професійного та автоматизованого перекладу, в даній роботі основна увага буде приділена методам навчання під наглядом.

Крім того, отримання даних можна розділити на дві основні задачі: перевірка і виявлення. У той час як перевірка намагається довести гіпотезу користувача, виявлення шукає ще невідомі закономірності в даних. Етап виявлення розділяється на опис, де система знаходить закономірності, щоб представити дані в зрозумілому форматі, і передбачення, де система намагається передбачити майбутні результати даних на основі закономірностей. Передбачення можна далі поділити на завдання класифікації і регресії. У той час як завдання класифікації мають фіксовані мітки, і кожен запис даних має одну з цих міток в якості значення атрибута класу, завдання регресії мають безперервні значення в якості вихідних даних. Дана робота присвячена алгоритмам, які намагаються передбачити, чи був даний технічний документ перекладений професійно або автоматично. Таким чином, завдання складається з двох фіксованих міток (професійний переклад і автоматичний переклад) і відноситься до сектору виявлення-передбачення отримання даних. На рисунку 1.1 показаний загальний вигляд таксономії отримання даних.

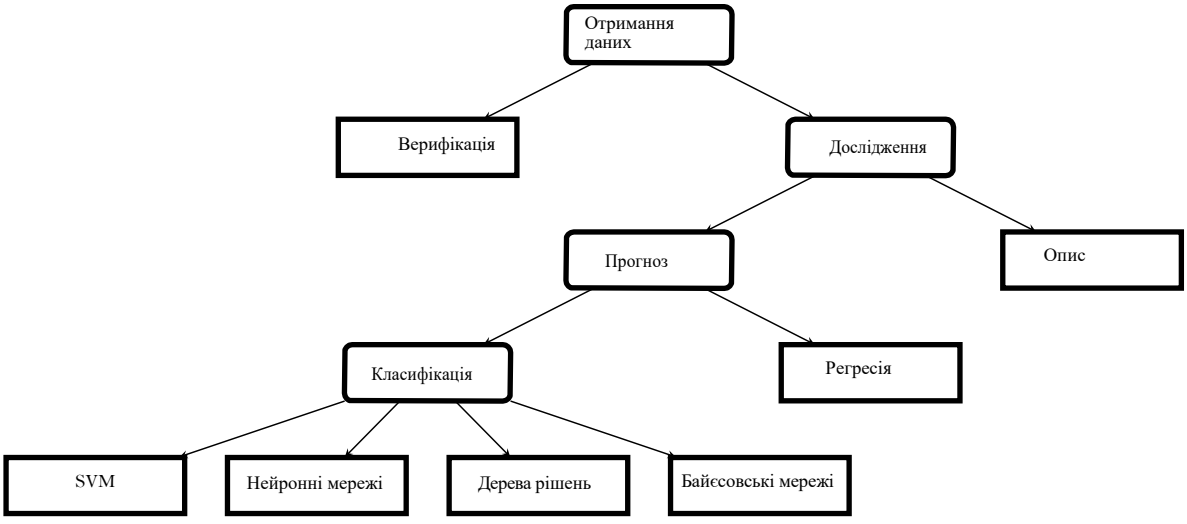


Рисунок 1.1 – Таксономія отримання даних

1.5 Методи машинного навчання

В контексті етапу отримання даних важливо вибрати правильний підхід до вирішення завдання. Для цього часто використовуються методи машинного навчання. Основна відмінність між людиною і комп'ютером вже давно полягає в тому, що людина схильна автоматично покращувати свій спосіб вирішення проблем. Люди вчаться на попередніх помилках і намагаються вирішити їх, виправляючи або шукаючи нові підходи до вирішення проблеми. Традиційні комп'ютерні програми не дивляться на результат виконання своїх завдань і тому не можуть поліпшити свою поведінку. Область машинного навчання вирішує саме цю проблему і передбачає створення комп'ютерних програм, які здатні навчатися і, отже, покращувати свої показники, збираючи більше даних і досвіду. У 1967 році перша програма розпізнавання образів змогла виявити патерни в даних, порівнюючи нові дані з відомими і знаходячи схожість між ними. З 1990-х років машинне навчання використовується в галузі отримання даних, адаптивних програмних систем, а також в області вивчення текстів і мов [11-13]. Як приклад: комп'ютерна програма, яка збирає дані про клієнтів магазину електронної комерції і створює з цих шматочків інформації більш персоналізовані рекламні оголошення, має здатність набувати нових знань і наближається до штучного інтелекту.

Крім того, системи машинного навчання зазвичай класифікуються за стратегією навчання, які часто визначаються кількістю висновків, які здатна зробити комп'ютерна програма:

1. Навчання за стратегіями описує стратегію, яку використовують всі традиційні комп'ютерні програми. Вони не роблять ніяких висновків, і всі їхні знання повинні бути прямо реалізуватись програмістом, оскільки програма не здатна робити будь-які висновки або перетворення з наданою інформацією.

2. Навчання на основі інструкцій охоплює всі комп'ютерні програми, які здатні здійснити перетворення інформації з заданої вхідної мови у вихідну мову. Незважаючи на те, що знання про те, як ефективно виконати це перетворення, все

ще даються програмістом, це вимагає невеликих форм виводів з боку комп'ютерної програми. Таким чином, це визначає окремий рівень системи навчання в порівнянні із стратегіями.

3. На відміну від навчання на основі інструкцій, навчання за аналогією намагається розвинути нові навички, які майже аналогічні існуючим і тому легко переймають, шляхом виконання перетворень відомої інформації. Ця система вимагає здатності створювати мутації і комбінації з динамічного набору знань. Вона створює нові функціональні можливості, які були невідомі вихіднній комп'ютерній програмі і тому вимагають великої кількості висновків.

Навчання на прикладах є одним з найбільш часто використовуваних видів навчання стратегії, оскільки воно забезпечує найбільшу гнучкість і дозволяє комп'ютерним програмам втілювати абсолютно невідомі навички або знаходити невідомі структури і патерни в даних. Навчання на прикладах - це техніка, яка часто використовується в задачах класифікації та отримання даних для передбачення класової мітки нових записів даних на основі динамічного набору відомих прикладів. У даній роботі запропоновані дослідження будуть пов'язані зі стратегіями і алгоритмами , що відносяться до цієї категорії.

Далі буде дано короткий опис найбільш поширених систем машинного навчання:

1.5.1 Дерево рішень

Дерево рішень - це метод класифікації, який орієнтований на легко зрозумілу форму подання і є одним з найбільш поширених методів навчання. Дерева рішень використовують набори даних, що складаються з векторів атрибутів, які в свою чергу містять набір класифікаційних ознак, що описують вектор, і атрибут класу, який зараховує запис даних до певного класу. Дерево рішень будується шляхом ітеративного розбиття набору даних по атрибуту, який якнайкраще розділяє дані на різні існуючі класи, поки не буде досягнутий певний критерій зупинки. Форма

подання дозволяє користувачам отримати швидкий огляд даних, оскільки Дерева рішень можна легко візуалізувати в деревовидному структурованому форматі, який легко зрозуміти людині.

Одним з перших алгоритмів навчання дерев рішень були ітеративний діхотомізатор 3 (ID3) і його наступник алгоритм C4.5 [14-16]. Ці алгоритми послужили основою для багатьох подальших розробок.

Дерева рішень - це орієнтовані дерева, які використовуються в якості інструменту підтримки прийняття рішень. Вони відображають правила прийняття рішень і ілюструють послідовні рішення.

В деревах рішень вузли можна розділити на кореневий вузол, внутрішні вузли та кінцеві вузли, також названі листям. Кореневий вузол являє собою початок процесу підтримки прийняття рішень і не має вхідних ребер. Внутрішні вузли мають рівно одне вхідне ребро і не менше двох вихідних ребер. Вони містять тест, заснований на атрибуті набору даних. Наприклад, такий тест може містити питання: "Чи є клієнт старше 35 років по атрибуту вік?". Вузли листя складаються з відповіді на проблему прийняття рішення, яка найчастіше представлена пророкуванням класу. Як приклад, проблемою прийняття рішення може бути питання про те, чи зробить клієнт інтернет-магазину покупку чи ні, при цьому класові передбачення можуть бути "так" і "ні". Листові вузли не мають вихідних і мають рівно одне вхідне ребро. Грані представляють рішення, прийняте в попередньому вузлі.

Якщо даний вузол *n*, то всі такі вузли, які відділені від *n* рівно одним ребром, називаються дочірні вузла *n*, а *n* називається батьком всіх своїх дочірніх вузлів. На рисунку 1.2 показаний приклад дерева рішень. Наприклад, запис даних, що має атрибути холодний , відмідь буде переданий вниз в ліве піддерево, оскільки його атрибут температури – холод, а потім вниз в лист "північ", який буде класифікований відповідною міткою.

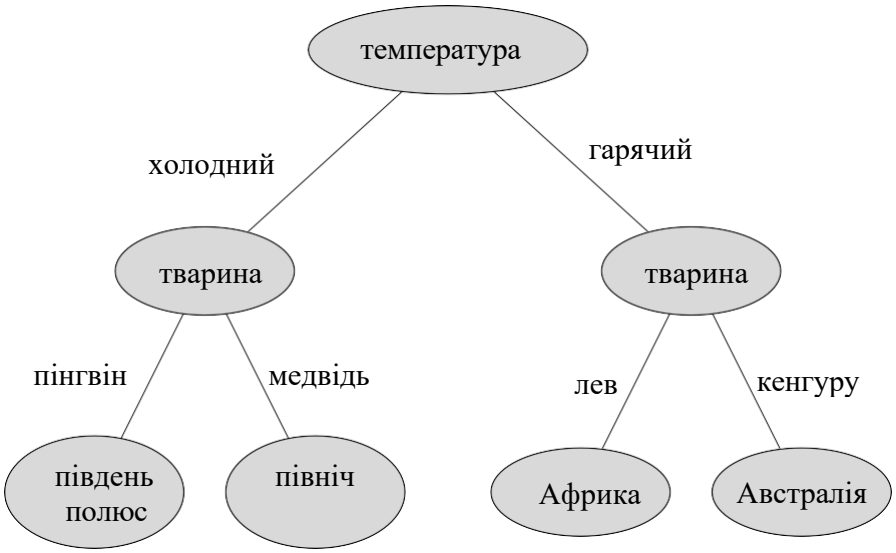


Рисунок 1.2 – Приклад дерева рішень

Навчання дерева рішень – це поширений метод отримання даних, який в основному використовується для класифікації. Його мета - передбачити значення цільового атрибута на основі ряду вхідних атрибутів. Навчання дерева рішень в контрольованому сценарії здійснюється за допомогою навчального набору для пошуку закономірностей в даних і побудови дерева рішень. Після цього набір раніше невидимих прикладів може бути використаний для прогнозування значення цільового атрибута. Навчальний набір містить записи даних в вигляді:

$$\left(\vec{x}, Y\right) = \left(x_1, x_2, x_3, ..., x_n, Y\right), \tag{1.1}$$

де Y – цільове значення атрибута,

x – вектор, що містить n вхідних значень

n – кількість атрибутів в наборі даних.

Для того щоб навчити дерево рішень і тим самим створити класифікатор, необхідний навчальний набір, що містить цільовий атрибут, вхідні атрибути, критерій поділу і критерій зупинки. В даному вузлі критерій поділу обчислює значення для всіх атрибутів. Це значення є мірою кількості інформації, одержуваної при розбитті вузла з даного атрибуту. Потім береться оптимальне значення з усіх атрибутів, і вузол розбивається на різні результати за відповідним атрибутом. На

цьому етапі процес пошуку найкращого розбиття серед атрибутів застосовується рекурсивно до всіх згенерованих піддерев до тих пір, поки не буде досягнуто критерію зупинки.

Загальними критеріями зупинки є:

– максимальна висота дерева досягнута;

– кількість записів у вузлі менше допустимого мінімуму;

– критерій найкращого поділу не подолав певний поріг з точки зору набраної інформації.

Якщо атрибут розділення має числовий тип, то немає можливості розділити записи на всі випадки атрибута. Це одна з основних переваг дерева рішень C4.5 в порівнянні з ID3. C4.5 також здатний обчислювати найкращі точки поділу для числових атрибутів і розділяти їх за допомогою операторів "більше, ніж" або "дорівнює" і "менше, ніж".

Навчання дерева рішень за допомогою цього автоматизованого процесу може привести до створення великих дерев рішень з ділянками дуже малої ваги з точки зору класифікації. Крім того, дерева мають тенденцію до надмірної підгонки, що означає, що вони занадто близько підходять до навчальних екземплярів. Це призводить до поганої продуктивності, коли ці дерева застосовуються до невидимих даних. Тому була розроблена техніка, названа обрізанням. Її мета - видалити з дерева рішень менш ефективні або непродуктивні частини, такі як частини, засновані на помилкових даних, або їх частині, які занадто добре підігнані. Це часто призводить до подальшого поліпшення точності і зменшення розміру дерева. Цей процес особливо важливий в зв'язку з тим, що кожен набір даних реального світу містить помилкові або зашумлені дані.

Тимчасова складність оригінального алгоритму вирощування дерев, що враховує тільки номінальні ознаки, становить $O(m * n2)$, де m – розмір навчального набору даних, а n – кількість атрибутів.

Найбільш трудомісткою частиною алгоритму вирощування дерева є обчислення приросту інформації для кожного атрибута. Для обчислення приросту

інформації необхідні значення відповідного атрибута для всіх записів даних в поточній навчальній підмножині. У гіршому випадку об'єднання усіх підмножин на всіх рівнях дерева рішень має той же розмір, що і вихідна навчальна множина. Таким чином, складність обчислення інформаційного виграшу для кожного рівня дерева вже становить O ($m * n$). Оскільки число рівнів дерева n, для найгіршого випадку загальна сумарна складність навчання дерева рішень становить O ($m * n^2$).

Після навчання дерева рішень, наступний процес полягає у використанні дерева для передбачення міток класів для невидимих записів даних. Для цього запис передається вниз від кореневого вузла до листа, перевіряючи відповідний атрибут в кожному вузлі і слідуючи по ребрах до відповідного листа. На рисунку 1.3 показаний процес навчання дерева рішень в псевдокоді, без урахування метричних ознак.

```
алгоритм Дерево рішень Процес навчання
1.  навчальна множина = S;
2.  набір атрибутів: A;
3.  цільової атрибут = C;
4.  критерій поділу = sC;
5.  критерій зупинки = зупинка;
6.  Gro(S, A, C, sC, stop)
7.  якщо stop(S) = false
8.  потім
9.  для всіх ai ∈ A
10.   знайти ai з найкращим sc (S);
11. позначити поточний вузол міткою a;
12. для всіх значень vi ∈ a
13.   помітити виходить ребро міткою vi
14.     Ssub= S, де a = vi;
15.     створити subNode = Gro (Ssub, A, C, sC, stop);
16. інакше currentNode = leaf;
17. помітити currentNode міткою ci, де ci- найбільш часто
зустрічається значення C ∈ S;
```

Рисунок 1.3 – Псевдокод навчання дерева рішень

Алгоритм починається з перевірки того, чи досягнуто критерій зупинки чи ні. Якщо так, то поточний вузол позначається найбільш загальним значенням з усіх існуючих міток класів для навчального набору. Якщо критерій зупинки не вірний, алгоритм обчислює значення розбиття для всіх атрибутів і позначає вузол атрибутом, відповідним найкращому значенню розбиття. Після цього він розбиває вузол на кілька вузлів, по одному для кожного значення обраного атрибута. Алгоритм викликає той же процес рекурсивно для всіх навчальних підмножин, що містять всі записи даних з відповідним значенням обраного атрибута.

1.5.2 Штучні нейронні мережі

Визначають штучні нейронні мережі як "математичну модель, яка заснована на біологічних нейронних мережах і тому є імітацією біологічної нейронної системи" [17-19]. У порівнянні зі звичайними алгоритмами, нейронні мережі можуть вирішувати досить складні завдання на значно більш простому рівні з точки зору складності алгоритму. Тому основною причиною використання штучних нейронних мереж є їх проста структура і система, що самоорганізується та дозволяє їм вирішувати широке коло завдань без додаткового втручання програміста. Наприклад, нейронна мережа може бути навчена на даних про поведінку покупців в інтернет-магазині і передбачити, чи буде робити людина покупку чи ні.

Штучна нейронна мережа складається з вузлів, так званих нейронів, зважених зв'язків між цими нейронами , які можуть бути адаптовані в процесі навчання мережі, і функції активації, яка визначає вихідне значення кожного вузла в залежності від його вхідних значень. Кожна нейронна мережа складається з різних шарів. Вхідний шар отримує інформацію з зовнішніх джерел, наприклад, значення атрибутів відповідного запису даних, вихідний шар виробляє вихідний сигнал мережі, а приховані шари з'єднують вхідний і вихідний шари один з одним. Вхідні значення кожного вузла в кожному шарі обчислюється сумою всіх вхідних вузлів,

помноженої на відповідну вагу зв'язку між вузлами. Крім того, нейронні мережі можна розділити на два основних типи:

1. Мережі з прямим зв'язком визначаються як всі мережі, які не отримують зворотного зв'язку від самої мережі. Це означає, що вхідні дані надходять в одному напрямку, від вхідних вузлів через від 0 до *n* прихованих вузлів до вихідних вузлів. Для переадаптації системи загалом немає ніякої інформації.

2. Рекурентні мережі визначаються як всі мережі, які містять опцію зворотного зв'язку і тому вони можуть повторно використовувати дані з більш пізніх етапів для процесу навчання на більш ранніх етапах.

Вихідне значення кожного вузла обчислюється шляхом використання всіх вхідних значень за заздалегідь визначеною функцією, яка однакова для кожного вузла мережі. Найбільш часто функцією є сігмоідна функція (o_j), яка визначається наступним чином:

$$o_j = \frac{1}{1 + e^{-i_j}}\ , \tag{1.2}$$

де i_j – сума вхідних вузлів j.

Двома основними перевагами цієї функції є її нормалізація до значень між 0 і 1 і її нелінійний характер, що призводить до більш швидкого навчання мережі та запобігання ефектів перевантаження і домінування. Ефект домінування виникає, коли один або кілька атрибутів роблять дуже великий вплив на прогнозований цільовий атрибут, роблячи інші атрибути безглуздими і, отже, домінуючи над ними.

На рисунку 1.4 показана нейронна мережа з прямолінійним рухом з трьома вхідними вузла у вхідному шарі, одним прихованим шаром і двома вихідними вузлами.

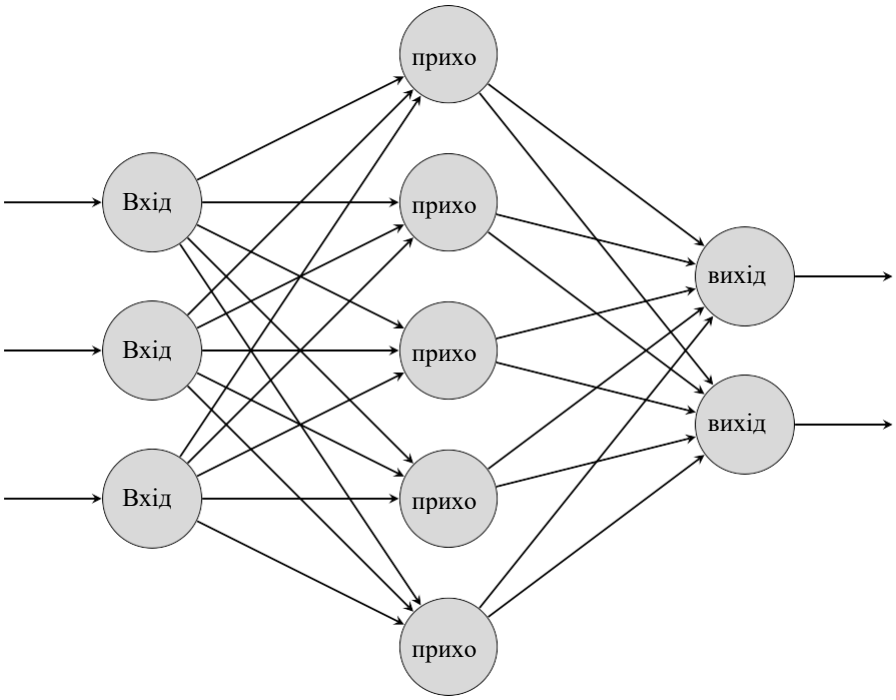


Рисунок 1.4 – Приклад штучної нейронної мережі

У сценарії контрольованого навчання штучні нейронні мережі можуть бути навчені за допомогою алгоритму зворотного поширення, який перебудовує ваги взаємозв'язків в нейронній мережі на основі локальних помилок.

1.5.3. Зворотне поширення в нейронних мережах

Зворотне поширення в нейронних мережах описує процес використання локальної помилки мережі для перенастроювання ваг взаємозв'язків в зворотному напрямку через нейронну мережу. В явному вигляді це означає, що після того, як був зроблений прогноз для набору вхідних значень, фактичне вихідне значення порівнюється з прогнозованим значенням і обчислюється помилка. Ця помилка потім використовується для перенастроювання ваг зв'язків, починаючи з країв, які безпосередньо з'єднані з вихідними вузлами мережі, і далі по мережі. Щоб навчити нейронну мережу, важливо розуміти основні параметри, які можуть бути використані для оптимізації процесу навчання:

1. Швидкість навчання визначає, як швидко відбувається процес навчання. Значення параметра лежить між 0 і 1 і множиться на локальну помилку для кожного вихідного значення. Тому, якщо коефіцієнт навчання дорівнює 0, то адаптація взагалі не відбувається. Правильна установка швидкості навчання має вирішальне значення для успіху процесу навчання. Якщо встановити занадто високе значення, ваги можуть коливатися і ускладнювати пошук оптимальних значень. Однак, якщо значення занадто мале, знайдені помилки не матимуть достатньої ваги, щоб підштовхнути мережу до нової оптимізації, і ваги можуть застрягти в локальних максимумах. Для того, щоб знайти правильні настройки, можна додати параметр загасання. Цей параметр забезпечує високе значення швидкості навчання на ранніх циклах процесу навчання, щоб уникнути застрявання в локальних максимумах, і примусово зменшує його в процесі навчання, щоб уникнути осциляції.

2. Ще один важливий параметр для нейронних мереж називається імпульсом. Він використовується для згладжування процесу оптимізації шляхом використання частки останньої зміни ваги і додати його до нової зміни ваги.

3. Мінімальна помилка є критерієм зупинки процесу навчання, аналогічним критерієм зупинки для дерев рішень. Як тільки сумарна помилка мережі падає нижче цього порога, процес навчання зупиняється.

Комбінуючи ці параметри, формула для обчислення нової ваги для з'єднання виглядає наступним чином:

$$W = l * \epsilon + m * W_p \ , \tag{1.3}$$

де W – нова зміна ваги;

l – швидкість навчання;

E – мінімальна помилка;

m – імпульс;

Wp – зміна ваги за попередній цикл.

1.5.4 Байєсовські мережі

Байєсовські мережі складаються з вузлів і спрямованих зв'язків між цими вузлами, які символізують залежності між ними. Вони являють собою ймовірні орієнтовані ациклічні графові моделі. Кожен вузол представляє атрибут, який представляє інтерес для даного завдання, наприклад, значення забруднення в містах для оцінки ймовірності розвитку раку легенів. Сама базова баєсова мережа називається Наївний Байес, і причина того, що вона називається Наївною, полягає в тому, що ця мережа передбачає відсутність залежностей між атрибутами. Цього майже ніколи не буває в практичних завданнях з пошуку даних, і тому цей метод, як правило, дає гірші результати, ніж більш детальні алгоритми. Звичайні байєсовські мережі використовують відомі дані для оцінки залежностей між атрибутами і міткою класу і використовують цю інформацію для розрахунку ймовірностей можливих різних результатів майбутніх подій [20-22]. Вони автоматично застосовують теорему Байеса до складних проблем і тому здатні отримувати знання про стан атрибутів і їх залежностях.

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \quad , \tag{1.4}$$

де *A* і B – це події;

P (X) – це ймовірність того, що подія X відбудеться;

P (X | Y) – це умовна ймовірність того, що подія X відбудеться, якщо відомо, що подія Y істинна.

Кожен вузол графа позначений розподілом ймовірності, яке визначає вплив батьківського вузла на дочірній вузол.

1.5.5 Навчання на основі прецендентів (kNN)

Навчання на основі примірників описує процес вирішення завдань на основі рішень аналогічних вже відомих задач, також відомий як навчання найближчих

сусідів [23-24]. Кожна система навчання на основі примірників вимагає набору параметрів:

1. Функція відстані, яка вимірює схожість між проблемами або записами даних. Це необхідно для вимірювання того, які з них є найближчими сусідами нової проблеми.
2. Ряд сусідів, які враховуються при вирішенні нової проблеми.
3. Вагова функція, яка дозволяє додатково оцінити знайдених сусідів для підвищення якості прогнозування та навчання.
4. Метод оцінки, який описує функцію, як використовуватиме знайдених сусідів для вирішення поставленого завдання.

Методи навчання на основі прецендентів відносяться до методів ледачого навчання, і це означає, що перед поданням запиту системі не було здійснено жодних обчислень над даними. Ці методи контрастують з методом навчання, таким як Дерева рішень, який намагається структурувати дані до отримання запитів.

1.5.6 SVM з підтримкою

SVM з підтримкою відносяться до області методів контрольованого навчання, тому для класифікації нових невидимих даних їм необхідні помічені, відомі дані. Основний підхід до класифікації даних починається зі спроби створити функцію, яка розбиває точки даних на відповідні мітки з (а) найменшою можливою кількістю помилок або (б) з найбільшим можливим запасом [25-26]. Це пов'язано з тим, що великі порожні області поруч з функцією розбиття призводять до меншої кількості помилок, оскільки мітки краще відрізняються один від одного.

Рисунок 1.5 демонструє, що набір даних цілком може бути розділено декількома функціями без будь-яких помилок. Тому маржа навколо розділяє функції, які використовуються як додатковий параметр для оцінки якості поділу. В даному випадку поділ А є найкращим, оскільки він більш точно розрізняє два класи.

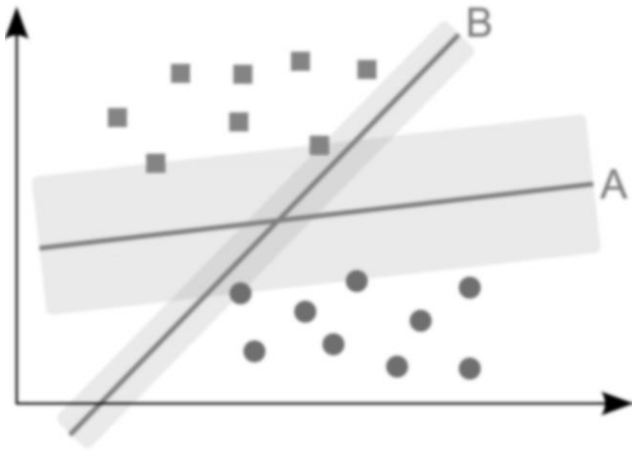


Рисунок 1.5 – Візуалізація роботи машини опорних векторів, що розділяє набір даних на два класи за допомогою двох різних лінійних роздільників, що призводить до різних розмірів полів навколо функцій поділу

Формально, опорні SVM створюють одну або кілька гіперплощин в *n*-вимірному просторі. Першою спробою в процесі розбиття даних завжди є спроба лінійно розділити дані на відповідні мітки. Наведений приклад, для завдання прогнозування ймовірності покупки клієнта в інтернет-магазині, використовує набір даних з *n* даними елементів, де кожна точка даних складається з мітки $y \in \{покупка, непокупка\}$ і вектор атрибутів *x*, що містить значення даних для цієї конкретної сесії. Тепер машина опорних векторів намагається знайти функцію, яка відокремлює всі точки даних (*x*, *y*) з *y* = *yes* від усіх елементів даних виду (*x*, *y*) з *y* = *no*. Якщо дані повністю разділимі лінійно, то отримана функція може бути використана для класифікації майбутніх подій. Штайнварт і Крістманн відзначають дві основні проблеми, пов'язані з цим підходом.

1. Дані можуть бути погано лінійно розділимі або взагалі не бути лінійно розділимі, що часто трапляється з реальними даними. У наведеному вище прикладі це може статися. Наприклад, два покупці поводяться абсолютно однаково в інтернет-магазині, але тільки один з них здійснює покупку. Це призведе до нероздільності даних, оскільки один і той же вектор атрибутів матиме різні мітки.

2. Друга проблема полягає в можливості переоцінки SVM. Щоб уникнути цього, дані повинні бути попередньо оброблені для виявлення шуму і прийняття

деяких неправильних класифікацій. Інше значення точності SVM будуть спотворені і приведуть до більш помилкової класифікації для майбутніх подій.

Перша проблема може бути вирішена за допомогою трюку з ядрами, відображаючи *n*-мірні вхідні дані в простір більш високої розмірності, де дані можуть бути розділені лінійно.

1.6 Постановка задачі

В даній роботі метод машинного навчання буде використаний в процесі виявлення знань для класифікації документів за типом перекладу (професійний переклад, автоматизований переклад). Далі буде запропоновано підхід до оцінки якості перекладених технічних документів. У зв'язку з цим вирішуємо два основних дослідницьких питання:

– як можна оцінити якість перекладу технічної документації, якщо є оригінал документа;

– як можна оцінити якість перекладу технічної документації, якщо оригінал документа недоступний.

Дана робота присвячена використанню методів і алгоритмів машинного навчання для оцінки перекладу технічної документації. В рамках даної роботи будуть вирішені два різні завдання. По-перше, переклади технічної документації будуть класифіковані і оцінені за допомогою алгоритму машинного навчання, що має доступ до оригіналу документа. У другій спробі алгоритм буде оптимізований для вирішення того ж завдання без знання оригіналу.

На основі вивчення існуючих методів і метрик буде запущений ітеративний процес виявлення знань для відповіді на поставлені дослідницькі питання. Цей процес включає в себе визначення критеріїв якості для переведених документів, впровадження необхідних метрик і алгоритмів, а також оптимізацію підходів машинного навчання для оптимального вирішення поставленого завдання. Важливо відзначити, що даний процес носить ітераційний характер, оскільки критерії та

ознаки, а також їх вплив на якість перекладу і можливості класифікації будуть визначатися шляхом оцінки результатів роботи алгоритмів з використанням бази даних технічних документів та їх перекладів. Використовуваний набір даних буде варіюватися від автоматизованих перекладів технічних документів за допомогою комп'ютерних систем переказу до ручних і професійних перекладів. Крім того, в ході цього ітераційного процесу використовувані методи і алгоритми будуть постійно змінюватися і оптимізуватися для досягнення найкращих результатів. Нарешті, процес і результати будуть піддані критичному аналізу, оцінки та порівняння між собою. Будуть вказані обмеження автоматизованого перекладу при сучасному стані техніки і дана перспектива для можливих подальших розробок і досліджень по цій темі.

У зв'язку з встановленням часових рамок, необхідно встановити деякі обмеження на дане дослідження, щоб робота була завершена в термін:

1. Класифікація та оцінка документів буде зосереджена на синтаксичних аспектах технічної документації, в той час, як семантичні аспекти залишаться осторонь.

2. У даній роботі особлива увага приділяється технічним документам. Ця орієнтованість має потенціал для неявного генерування знань в процесі машинного навчання, завдяки меншому розміру словникового запасу в порівнянні з відсутністю обмежень на текстові домени. Інші домени документів не будуть розглядатися в даній роботі.

3. Оскільки в розглянутих технічних документах не було багаторазових професійних перекладів, для оцінки технічних документів не будуть використані рекомендації людей. Замість цього будуть використовуватися псевдопосилання, щоб обійти відсутність людського перекладу.

4. Ця робота зосереджена на оцінках, заснованих на результатах застосування методів машинного навчання. Інші методи, такі як створення нової метрики, яку можна порівняти з метриками АДО або Meteor, не братимуться до уваги.

Висновки до розділу

Робота представляє особливий інтерес для досліджень в області машинного перекладу і оцінки машинного перекладу за допомогою комбінацій різних метрик машинного перекладу і підходів машинного навчання. Групою інтересів для даного дослідження є міжнародні компанії, особливо орієнтовані на експорт, в зв'язку з тим, що знайти технічно грамотних перекладачів при обмеженому бюджеті явно проблематично. По-друге, ця робота представляє інтерес для всіх видів клієнтів, оскільки вона може призвести до довгострокового поліпшення доступної інформації по певних продуктах. Це може представляти додатковий інтерес для клієнтів і компаній, що працюють з мало розповсюдженими мовами.

У наступному розділі будуть розглянуті теоретичні передумови даної роботи з упором на процес виявлення знань, підходи машинного навчання, машинну трансляцію і технічну документацію. Після опису взятої методології емпіричні дані і відповідні результати експериментів будуть показані у відповідному розділі. Обговорюватимуться отримані результати, критично розглянуті використані підходи і методи, а також вивчена обґрунтованість і надійність представлених результатів. Нарешті, будуть підведені підсумки роботи та надані перспективи можливих майбутніх досліджень і розширення даної роботи.

Розділ 2
Розробка моделі оцінки методу класифікації

2.1 Оцінка якості машинного навчання

Дуже важливою частиною машинного навчання є проблема того, як комп'ютерна програма помічає, які з її результатів були слушними, а які містили помилки. Прикладом, коли це не створює проблем для алгоритму, може бути комп'ютерна програма, яка намагається передбачити, чи зробить клієнт в магазині електронної комерції покупку чи ні. Після введення даних в журнал буде занесена інформація про те, чи купив покупець товар чи ні, яка потім може бути використана для оцінки роботи алгоритму. Більш складні сценарії виникають в тих областях досліджень, де доступ до реальних даних обмежений або відсутній, наприклад, при оцінці перекладу документів. Це вимагає додаткових зусиль людини для ранжирування заданих перекладів по класах, щоб в результаті можна було порівняти результати роботи комп'ютерної програми. Оцінка завдань класифікації зазвичай проводиться шляхом розбиття набору даних на навчальний і тестовий. Потім алгоритм машинного навчання навчається на першому, а тестовий набір даних використовується для розрахунку показників ефективності, щоб оцінити якість алгоритму. Загальна проблема алгоритмів машинного навчання полягає в обмеженому доступі до тестових і навчальних даних. Тому перебір може стати серйозною проблемою при оцінці цих програм. Для вирішення цієї проблеми поширеного підходу є використання Fold Cross Validation. Крос-валідація описує процес поділу всього набору даних на X частин і послідовного використання кожної з них в якості тестового набору даних, в той час як інші об'єднуються в навчальні дані. Після цього показники ефективності усереднюються по всіх процесах валідації. Не існує ідеального показника для кожного суб'єкта оцінки алгоритмів машинного навчання, оскільки кожен має свої недоліки і переваги. Найбільш важливими факторами для оцінки ефективності програми машинного навчання є наступні:

1. Показник помилкової класифікації описує відносну кількість помилково класифікованих даних в наборі даних. Якщо y_i - пророкування для точки даних *i*, а y_i - фактична мітка, то помилкова класифікація визначається як

$$misc_n = \frac{1}{n} * \sum_i (y_i \neq \hat{y}_i) \tag{2.1}$$

Щоб уникнути цієї проблеми, використовується бенчмаркінг.

2. Бенчмаркінг описує процес порівняння значення показника з еталонним значенням, щоб посилити його твердження. Наведений приклад, для бінарної класифікації. У сценарії контрольованого навчання еталонним алгоритмом може бути класифікатор, який завжди передбачає найбільш поширений клас. Рівень помилкової класифікації може бути описаний по відношенню до еталону. У цьому випадку рівень помилкової класифікації в 20% при рівному розподілі класів дасть відносне поліпшення в порівнянні з еталоном в 30 пунктів.

3.Значення точності, також назване позитивним значенням передбачення, визначається як відносна кількість правильно класифікованих примірників серед всіх правильно класифікованих примірників. Це можна проілюструвати на прикладі електронної комерції, згаданої раніше. Значення точності 1 означає, що кожен клієнт, класифікований міткою “продаж”, дійсно робить покупку. Однак важливо відзначити, що це не впливає на кількість клієнтів, класифікованих як "без покупки", які також здійснюють покупки.

5. Значення повнота, також назване чутливістю, визначається як відносна кількість справжніх класифікованих примірників серед всіх справжніх примірників. Для запропонованого прикладу, це означає, що значення повнота, рівне 1, означає, що кожен покупець був класифікований як покупець. Важливо відзначити, що значення повнота, рівне 1, може бути легко досягнуто шляхом класифікації кожного примірника набору даних, як покупки.

6. F-міра покликана об'єднати показники повнота і точність, використовуючи

середнє гармонійне між ними:

$$F_1 = 2 * \frac{precision * recall}{precision + recall} \quad (2.2)$$

7. Одним з кращих підходів для ілюстрації ефективності програм машинного навчання є матриця змішування, також звана таблицею випадковостей, яка розрізняє між істинно позитивними, помилково позитивними, істинно негативними і помилково негативними прогнозами.

істинний позитив	хибне заперечення
хибне спрацьовування	справжній негатив

Рисунок 2.1 – Приклад матриці помилок

Основна проблема неправильної класифікації полягає в тому, що її результати сильно залежать від кількості міток або розподілу даних між мітками класів. Наведений приклад досягнення коефіцієнта помилкової класифікації 0,03 може виглядати дуже багатообіцяючим без подальшого контексту, але в прикладі, де 97% набору даних позначені класом a і 3% - класом b, цього неважко досягти. Те ж саме відноситься і до відмінностей в кількості доступних класів. Показник помилкової класифікації в 20% або нижче символізує явно кращу систему машинного навчання для набору даних з трьох класів, ніж для одного.

Основний недолік матриці помилок полягає в тому, що вона вимагає інтерпретації людиною. У даній роботі методи машинного навчання будуть використовуватися в контексті проблем машинного перекладу.

2.2 Ефективність машинного перекладу

Тема подолання мовних бар'єрів за допомогою пристроїв має довгу історію, починаючи з 17 століття. У той час виникла ідея створення універсальної мови для полегшення взаєморозуміння між людьми з усього світу. Приклади в той час складалися з символів і логічних принципів. Однак в середині 20 століття виникла тема автоматизації процесу перекладу, а в 1952 році почалися перші конференції з машинного перекладу (МП) для обміну знання та дослідження в цій специфічній галузі. Сьогодні машинний переклад визначається як переклад з однієї природної мови (вихідна мова) на іншу мову (мову перекладу) за допомогою комп'ютеризованих систем, за допомогою людини або без неї [27-28].

Машинний переклад включає в себе такі області, як машинний переклад за допомогою людини і машинний переклад за допомогою людини, які визначаються як виробництво перекладів системами за допомогою перекладачів-людей. При цьому комп'ютеризовані системи, що надають словники або інші засоби допомоги вищого рівня перекладачам, не включаються. При сучасному рівні розвитку техніки, машинний переклад все ще далекий від того, щоб ідеально перевести будь-який текст з будь-якої початкової мови на будь-яку мову перекладу. Проте, після більш ніж п'яти десятиліть розробок в цій області, комп'ютеризовані системи здатні створювати "сирі" переклади, які зазвичай все ще перевіряються людьми для підтвердження їх достовірності. Крім того, такі машини часто орієнтовані на певну область, лексику або тип тексту для подальшого поліпшення якості перекладу. Сьогодні існує два різних підходи до машинного перекладу, які будуть описані нижче. Переклад, заснований на правилах, охоплює всі процеси перекладу, які засновані на правилах, що стосуються синтаксичних, семантичних і прямих

словесних аспектів тексту. Використання перекладу на основі правил - це компроміс між складністю і якістю перекладу. Це пов'язано з тим, що однією з основних переваг систем, заснованих на правилах, є відсутність межі якості для переказів. В теорії кожна помилка може бути виправлена шляхом впровадження правила для даного конкретного випадку, а оскільки кількість використовуваних правил необмежена, кожна помилка може бути виправлена. Це теоретична точка зору, оскільки кількість можливих комбінацій слів або пропозицій і їх різних значень, як правило, занадто велика. Тому при практичному застосуванні слід очікувати певної кількості помилок.

Прямий переклад був першим типом машинного перекладу, і він описує всі системи, які розробляються для одного конкретного напрямку перекладу. Це означає що системи здатні перекладати тексти тільки з певної вихідного мови на одну конкретну мову перекладу, що зазвичай робиться за допомогою словникового перекладачу без спроби зрозуміти контекст або сенс пропозиції або тексту. Морфологічний аналіз, такий як визначення закінчення слова і відмінювання, проводиться в дуже малому ступені. На рисунку показаний загальний процес прямого машинного перекладу.

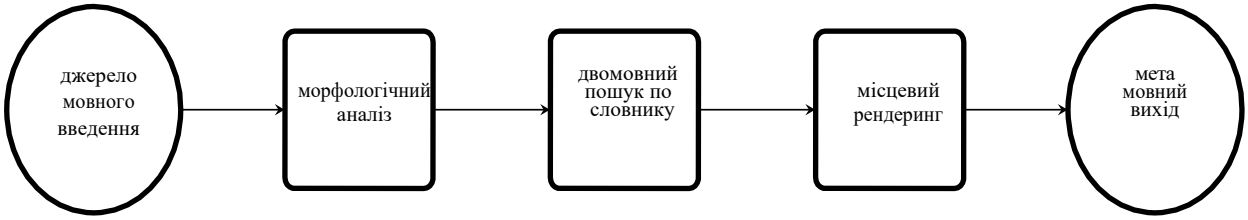


Рисунок 2.2 – Процес прямого перекладу

Через серйозний недолік розуміння синтаксису і контексту в методі прямого перекладу були прийняті нові підходи до вирішення проблеми машинного перекладу.

Другий тип перекладу на основі правил називається перехідний переклад, оскільки він використовує етап перекладу для аналізу вихідної і цільової мови і виявлення зв'язків між словами і можливих значень для створення більш якісного перекладу.

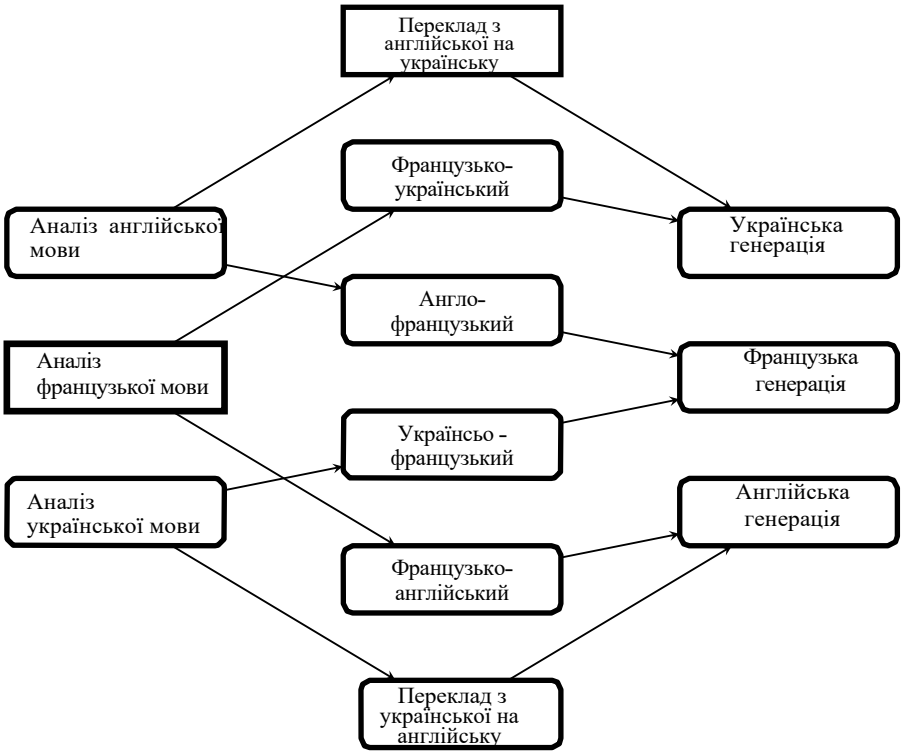


Рисунок 2.3 – Візуалізація залежності мови та напрямки перекладу на етапі перекладу в процесі трансферного перекладу

Для виявлення залежностей і кореляцій слів використовується метод маркування частин мови, при якому кожне слово маркується відповідною частиною мови наприклад, дієслово, іменник, прикметник або прислівник. Сьогодні підходи, засновані на перенесенні, є найбільш часто використовуваним методом, завдяки своїм явним перевагам перед прямим перенесенням і порівняно простій структурі. Один з основних недоліків структури перенесення - залежність від мови. Для підтримки декількох мов і двонапрямлених завдань необхідна велика кількість процесів переносу. Зокрема, для того щоб додати в систему $n + 1$ мову, потрібне

додаткова кількість процесів переносу. Необхідно 2*n кроків перенесення. Це призвело до розробки третього типу перекладів, заснованих на правилах, наприклад, переклад на інтерлінгва.

Переклад на інтерлінгва був другим підходом до вирішення проблеми відсутності синтаксичного і семантичного розуміння підходу прямого перекладу. Основна відмінність від трансферного перекладу полягає у використанні міжнародної допоміжної мови в якості окремого кроку в процесі перекладу. Використання такої допоміжної мови, яка є сумісною з дуже широким спектром різних природних мов, дозволить різко скоротити кількість необхідних етапів перекладу, оскільки необхідні етапи трансформації полягають у перетворенні з кожної вихідної мови на допоміжну і навпаки. У підсумку, підтримка в цілому n мов призводить до необхідності 2n кроків перекладу. Додавання мови продукує два додаткових кроки, крок аналізу для створення проекції з нової мови на допоміжну мову і крок генерації для створення речень на нову мову. Це контрастує з 2n додатковими кроками в процесі трансферного перекладу. На рисунку представлений процес перекладу на інтерлінгва для двох мов. Основна проблема, пов'язана з цим підходом, полягає в тому, що створення такої допоміжної мови не є тривіальним, а підтримка кожної існуючої мови практично неможлива. Крім того, оскільки генерація пропозиції на цільову мову строго незалежна від вихідної пропозиції, неможливо оптимізувати аналіз вихідного тексту відповідно до цільової мови. Це накладає важкі вимоги на допоміжну мову, оскільки він повинен надавати всі можливі аналізовані аспекти вихідного тексту, так як вони можуть знадобитися на етапі генерації. Щоб надати кожну частину інформації, допоміжна мова повинна розуміти сенс пропозиції. Це перетворює проблему перекладу в задачу обробки природної мови на семантичному рівні, яка все ще перебуває на ранніх стадіях дослідження.

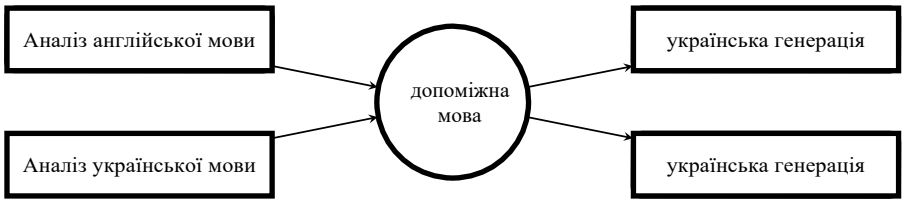


Рисунок 2.4 – Процес перекладу на інтерлінгва для двох підтримуваних мов

Відповідним способом візуалізації різних підходів до машинного перекладу на основі правил є використання піраміди, як показано на рисунку.

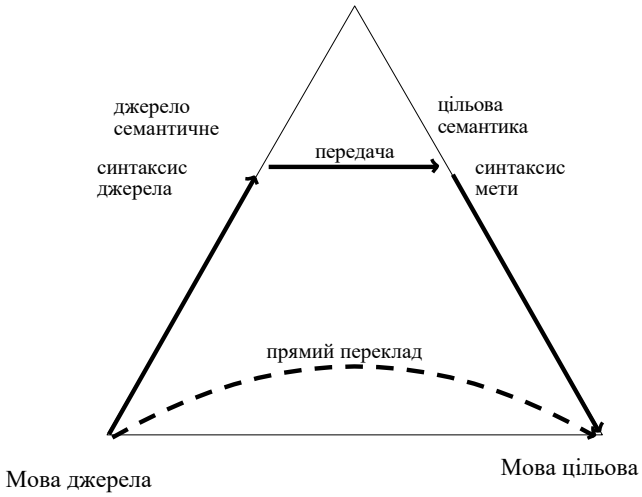


Рисунок 2.5 – Візуалізація обсягу аналізу, виконуваного в процесі машинного перекладу на основі правил

Процес перекладу починається в лівому нижньому кутку трикутника. Залежно від підходу до перекладу виконується певний обсяг аналізу, перш ніж буде виконано етап перекладу для створення вихідної пропозиції з використанням аналізу мови перекладу. Оскільки при прямому перекладі аналіз практично не використовується, переклад здійснюється безпосередньо з вихідної мови на мову перекладу. З іншого боку, при підході Iterlingua виконується максимально можливий аналіз для перетворення вхідного тексту в нейтральну незалежну форму і використання цієї форми для створення вихідного тексту.

2.3 Машинний переклад на основі прикладів

Друга велика область типів машинного перекладу називається машинний переклад на основі прикладів (МПОП). Існує безліч підходів до машинного перекладу, які відносяться до галузі перекладу на основі прикладів, але основною визначальною частиною цієї області є використання бази даних, що містить вже перекладені пропозиції або частини тексту. Процес перекладу в системах МПОП складається з порівняння фрагментів нових, неперекладених частин тексту з базою даних, співставлення їх з відповідними аналогічними фрагментами тексту й об'єднання їх в новий перекладений текст [27-28]. Співставлення нових елементів даних з найбільш схожими в базі даних в основному здійснюється шляхом обчислення міри відстані між фрагментами і вибору "найближчого" серед прикладів. Для того щоб вказати на схожість між традиційним (заснованим на правилах) машинним перекладом і МПОП, Сомерс пропонує адаптовану версію трикутника, як показано на рисунку 2.6. Частина співставлення замінює етап аналізу, вибір відповідних фрагментів тексту мовою перекладу замінює етап перенесення, і замість генерації нового речення на основі результатів аналізу ці фрагменти об'єднуються.

Основною проблемою, з якою стикаються системи МПОП, є необхідність наявності відповідної бази даних, що містить двомовний набір прикладів. Сомерс зазначає, що вибір корпусу, орієнтованого, наприклад, на певну область, може значно поліпшити якість перекладу. Крім того, основними моментами, що викликають заклопотаність щодо набору прикладів, є довжина фрагмента тексту і розмір набору даних. Вибір довжини прикладу - це завжди компроміс. Чим довше запис даних, тим нижча ймовірність знайти збіг, що робить запис даних не зручним у використанні, але чим коротше запис, тим вище ризик помилкового віднесення її до не зовсім подібного фрагмента.

Рисунок 2.6 – Адаптована версія трикутника

Найбільш часто використовувана довжина фрагмента - це довжина речення. Це не обов'язково пов'язано з тим, що така довжина є найкращою, а скоріше тому, що закінчення речення легко визначити і розбити текст на фрагменти. Розмір набору прикладів - простіша проблема. Зазвичай вважається, що більша кількість прикладів призводить до кращих результатів, хоча, як згадувалося вище, зосередженість на конкретній області може бути перевагою навіть у порівнянні з великим набором прикладів. Крім того, передбачається, що в якийсь момент користь від додавання нових прикладів в набір даних зникає.

2.4 Статистичний підхід

Підхід статистичного машинного навчання є частиною МПОП, оскільки використовує набір даних вже перекладених фрагментів. Відмінність від інших методів МПОП полягає в тому, що статистичні підходи намагаються знайти закономірності в наборі даних, обчислити ймовірності і зробити розумні висновки на основі отриманих результатів.

З появою області машинного перекладу виникла і проблема оцінки систем МП. Без належного підходу до визначення якості роботи машини перекладу

подальші дослідження не мали сенсу. Перші спроби оцінки були засновані на людських, ручних метриках. Першими двома найбільш поширеними підходами були оцінка швидкості і адекватності. Швидкість вимірюється людиною, яка вільно володіє мовою перекладу, і дозволяє визначити, читається чи речення перекладу, без урахування правильності перекладу. Адекватність, з іншого боку, оцінює, чи була перекладена основна інформація без урахування граматичної правильності чи швидкості.

Обидва показники зазвичай вимірюються за шкалою від п'яти до семи балів. Незважаючи на те, що неодноразово було доведено, що ці метрики не є достатні і не мають високу кореляцію, людська оцінка як і раніше використовується як еталон або базова лінія для оцінки інших метрик автоматизованого перекладу.

При оцінці людиною швидкості і адекватності перекладу виникають в основному дві проблеми.

Перша полягає в тому, що для оцінки перекладу вимагається велика кількість людської допомоги, що робить безглуздим сенс автоматичного перекладу текстів. По-друге, оцінка тексту може сильно відрізнятися у двох різних оцінюючих осіб.

Для вирішення цих проблем в кінці 20-го століття почалася розробка математичних і автоматизованих метрик. Ці автоматизовані процеси зазвичай полягають в порівнянні результатів перекладу з набором еталонних перекладів і обчисленні відстані або міри схожості між ними. Далі наводимо огляд існуючих в даний час алгоритмів для найбільш важливих оцінок перекладу.

2.5 Переклад в обидві сторони

Метод перекладу туди-назад (ПТН) був одним з перших методів, використаних для оцінки якості систем машинного перекладу. Для того щоб оцінити роботу програми перекладу, їй давався фрагмент тексту, який спочатку переводився на іншу мову, а потім назад на вихідну. Отримані результати можна було легко порівняти з вихідним фрагментом тексту. Широко поширена думка, що ПТН-

переклад є невідповідним підходом для оцінки якості речення[27]. Це пов'язано з тим, що переклад в обидві сторони вимагає наявності двох добре працюючих систем перекладу (з оригіналу на іноземну мову і навпаки) і, як наслідок, зазвичай призводить до великої кількості помилок. Крім того, якщо в підсумковому документі не буде помилок, це може означати, що помилки виникли під час першої половини процесу переведення туди-назад і були виправлені на зворотному шляху. Тому в ході досліджень було розроблено безліч додаткових метрик для оцінки якості перекладу, які будуть розглянуті в наступних підрозділах.

Коефіцієнт помилок в словах (ПС) був однією з перших метрик, які використовуються для оцінки перекладів, і до сих пір є стандартною методикою оцінки автоматичного розпізнавання мови. Ключовим аспектом коефіцієнта помилок в словах є використання відстані Левенштейна як міри схожості між двома реченнями . Це робиться шляхом визначення, які слова не можуть бути знайдені в новому перекладі (видалення (D)), які слова не можуть бути знайдені в еталонному перекладі (вставки (I)), які слова були замінені іншими (заміни (S)) і які слова відповідають один одному. Всі ці правки розраховуються по відношенню до існуючого професійного перекладу того ж фрагмента тексту, який називається еталонним перекладом. Коефіцієнт помилок в словах розраховується як сума замін, вставок і вилучень, поділена на кількість слів у даному фрагменті тексту (N):

$$\text{ПС} = \frac{S+I*D}{\text{N}} \tag{2.3}$$

Коефіцієнт помилок в словах може бути розширений для використання в перекладах з декількома посиланнями (МПС) шляхом обчислення коефіцієнтів помилок у всіх окремих словах і використання найближчого посилання в якості результату. Однак такий підхід спрямований тільки на включення ПС в довідкове завдання, оскільки обчислений результат не отримує ніякої додаткової інформації при використанні декількох посилань. Основний недолік коефіцієнта помилок в словах для перекладу тексту полягає в тому, що може існувати кілька правильних

перекладів одного і того ж речення, в яких не використовується ні однаковий порядок слів, ні навіть однакові слова, що в обох випадках призведе до гірших показників ПС.

2.6 Коефіцієнт помилок перекладу

Коефіцієнт помилок перекладу, також названий коефіцієнтом редагування перекладу (КРП), був розроблений в 2005 році. Він використовувався для розрахунку кількості правок, необхідних для перетворення автоматизованого перекладу в правильний фрагмент тексту з точки зору його швидкості і адекватності. Щоб нормалізувати метрику КРП, скільки разів редагували переклад в контекст довжини тексту посилання або, в разі кількох посилань, середньої довжини посилань. Отримана формула для оцінки КРП визначається наступним чином:

$$\text{TEP} = \frac{\text{скільки разів редагували}}{\text{середня кількість слів у посиланнях}} \tag{2.4}$$

Важливо відзначити, що розділові знаки вважаються повними словами для оцінки КРП, а наступні дії вважаються правками:

- вставка слів, відсутніх в перекладеному фрагменті тексту;
- видалення слів, що не існують в довідковому тексті;
- заміна слів в перекладеному тексті;
- зрушення від *1* до *n* послідовних слів на *0 - n* кроків в будь-якому напрямку у фрагменті.

Обчислення мінімальних відстаней редагування є NP-повною завданням, і її рідко можна наближено вирішити за допомогою пошуку або динамічного програмування.

2.7 Алгоритм двомовної оцінки

Алгоритм двомовної оцінки (АДО) був представлений у 2002 році. Його мета - створити швидкий і надійний метод оцінки якості машинного перекладу без активної участі людини. Визначають якість як близькість тексту-кандидата до професійно переведеними людиною посиланнях. АДО забезпечує високу кореляцію між автоматичними і людськими оцінками тексту і тому є основним алгоритмом для оцінки та поліпшення систем машинного перекладу. Алгоритм не залежить від мови та в основному використовується для оцінки повних документів через недоліки оцінок на рівні речення.

Метрика АДО заснована на вимірі точності, підрахунку кількості однакових слів у перекладі-кандидата і еталонному перекладі, перекладеному людиною. Після цього збігу діляться на загальну кількість слів, використаних в перекладі кандидата.

У наступному прикладі показано пропозицію-кандидат, який отримав оцінку 5/6, що свідчить про хороший переклад.

Кандидат 1: Птах сидить на дереві.

Довідка 1: На дереві сидить птах.

Довідка 2: Птах сидить на дереві.

Для вирішення проблеми, пов'язаної з тим, що низькоякісні речення-кандидати, дають більш високу оцінку (в даному випадку 6/6), ввели модифіковану точність уніграми, яка обмежує оцінку одного слова, вважаючи опорне слово вичерпаним після визначення збіжності слова-кандидата. Таким чином, Modified Unigram Точність становить 1/6.

Щоб врахувати швидкість перекладу, модифікована точність розраховується для блоків слів, так званих n-грам , де n - кількість послідовних слів в блоці фрагмента тексту.

Щоб врахувати різну довжину перекладів кандидатів і посилань і переконатися в тому, що дуже короткі кандидати не дають порівняно високих результатів, до модифікованої точності N-грам додається штраф за стисліть.

Оскільки переклади-кандидати, які довші еталонних, вже караються по Modified N-gram Точність, неявний коефіцієнт штрафу для довгих речень дорівнює 1. Більш короткі перекази караються штрафом за стисліcть (ШС). Повна формула для метрики АДО на рівні корпусу:

$$\text{АДО} = ШС * (\sum_{i=1}^{n} w_i log p_i), \quad (2.5)$$

де

$$\text{ШС} = \begin{cases} 1 \; if \; c > r \\ e^{1-\frac{r}{c}} \; if \; c \leq r \end{cases};$$

w_i – це позитивні ваги для кожного використовуваного набору n-грам, які обчислюються наступним чином використовуючи середнє геометричне значення збігів по n;

n – максимальна довжина n-грами;

i – n-грами , довжина блоку;

p_i – модифікована точність n-грамми;

c – загальна довжина корпусу перекладів-кандидатів;

r – загальна довжина найбільш підходящого еталонного корпусу.

Представлена метрика АДО поєднує в собі швидкість обчислень і незалежність від мови з високою кореляцією людських суджень і тому широко використовується для машинного перекладу.

Однак алгоритм ігнорує деякі специфічні характеристики мови, такі як робота з синонімами, і тому найбільш продуктивний на рівні документа. Крім того, до уваги береться важливість і інформаційна цінність оцінюваного тексту, а окремі слова і фрази переоцінюються. Такі варіанти, як метрика NIST або оцінка Meteor, намагаються усунути ці недоліки і будуть розглянуті в наступних підрозділах.

2.8 Розширення алгоритму двомовної оцінки - NIST

Метрика NIST названа в честь Національного інституту стандартів і технологій США. Вона розширює базову метрику АДО і враховує деякі недоліки. Вона вводить зважування n-грам відповідно до їх інформаційну цінність і тому відносно знижує вагу загальних n-грам. Вона також замінює середнє геометричне значення числа спільних входжень по N, що використовується в алгоритмі АДО (N - кількість слів у даному тексті), на середнє арифметичне число n-грам. Формула NIST виглядає наступним чином:

$$NIST = \sum_{i=1}^{n} \left(\frac{\sum info(\text{спільно зустрічаються } \mathbf{n}-\text{грами })}{\text{Кількість } n-\text{грам в кандидата}} \right) \quad (2.6)$$

Info (x) – індивідуальний інформаційний вага n-грами для n-грами x, яка визначається наступним чином:

$$Info(w_1, w_2, \dots, w_n) = log_2 \left(\frac{\text{кількість входжень } w_1,\dots,w_{n-1}}{\text{кількість входжень } w_1,\dots,w_n} \right) \quad (2.7)$$

Метрика NIST є коригуванням раніше описаної оцінки АДО і також орієнтована на оцінку документів, а не на оцінку рівня речень. У порівнянні з АДО, NIST дозволяє більш точно оцінювати переклад і вважає за краще рідкісні n-грами звичайним, виходячи з того, що рідкісні n-грами містять більше інформації, ніж звичайні.

Хоча NIST досягає кращих результатів у порівнянні з АДО, коли мова йде про подібність з людськими оцінками перекладів, для розрахунку ваги інформації n-грам потрібно кілька посилань.

2.9 Метрика для оцінки перекладу з явним упорядкуванням

Meteor (Metric for Evaluation of Translation with Explicit Ordering - Метрика для оцінки перекладу з явним упорядкуванням) - це ще один алгоритм оцінки машинного перекладу, заснований на АДО. Метрика була розроблена для усунення основних проблем алгоритму АДО і дозволяє досягти кращих результатів на рівні речень.

Meteor заснований на середньому гармонійному значенні точності і повнота уніграмм, причому Повнота має більшу вагу, ніж точність.

Спочатку алгоритм створює вирівнювання між двома реченнями, кандидатом і еталонним реченням. Вирівнювання є відображення на основі уніграмм і намагається пов'язати кожну уніграмму кандидата з відповідною уніграммою еталона.

Кожна уніграмма з кандидата може відповідати максимум одному опорному слову. Meteor розглядає чотири різних види збігів:

– точний: відповідні слова ідентичні, вони мають одну і ту ж схему перекладу;

– кореневий: вирівняні слова мають загальний корінь слова і тому збігаються;

– синонім: значення слів однакові, вони є синонімами один одного;

– перефразування: повні перекази вважаються співпадаючи , якщо вони мають однаковий сенс.

Для остаточного вирівнювання кількість покритих слів у реченнях максимізується, а кількість фрагментів мінімізується. Частини визначаються як ланцюжок збігів, які є безперервними і однаково впорядкованими в порівнюваних пропозиціях. Вихідна установка перетворюється в формулу, яка використовує середнє гармонійне, яке розраховується за допомогою одногрaмових точностей і відхилення. Одногрaмова точність P розраховується як:

$$P = \frac{m}{w_t} \tag{2.8}$$

де m являє собою кількість збігів, а w_t – кількість слів у тексті-кандидаті.

Повнота уніграмм R обчислюється як:

$$P = \frac{m}{w_r}, \tag{2.9}$$

де m залишається незмінним, а w_r – кількість слів у посиланні.

Обидві оцінки об'єднуються до середнього гармонійного значення наступним чином:

$$Fmean = \frac{P*R}{\alpha P+(1-\alpha)*R} \tag{2.10}$$

Для того щоб штрафувати за короткі речення і множинні фрагменти, Meteor вводить штраф p, який вираховується як:

$$\rho = \gamma * \left(\frac{c^{\beta}}{u_m}\right), \tag{2.11}$$

де c – кількість ланцюжків, а u_m - кількість зіставлених уніграмм.

Підсумкова оцінка Метеора складається з $Fmean$ і штрафу p в наступному вигляді

$$Meteor = Fmean * (1 - p) \tag{2.12}$$

Для настройки результатів Meteor, зважують уніграми відповідно до їх збігу в цільовій мові і рекомендують налаштовувати параметри α, β і γ. Загальноприйнятими значення є:

– α = 0.9;

– β = 3;

– γ = 0.5.

Основною перевагою метрики Meteor є використання таблиць перефразування і синонімів, а значить, можливість розпізнавати подібності на невеликих обсягах тексту з більш високою точністю.

Однак через використання великої кількості зовнішніх даних, таких як таблиці перефразування і синонімів, а також обширного методу пошуку збігів в цілому, метрика вимагає не тільки мовної підготовки, а також великої кількості обчислювального часу для розрахунку в порівнянні з показником АДО.

Крім того, якщо слідувати реченню щодо оптимізації альфи, бети і гами окремо, це ще більше збільшить трудовитрати.

Проте, не дивлячись на великі витрати на підготовку і обчислення, Meteor в даний час визнана кращою метрикою для оцінки рівня речень.

У даній роботі методи машинного навчання використовуються в контексті проблем машинного перекладу з акцентом на специфічну область технічної документації.

2.10 Технічна документація

Технічна документація - це загальний термін для кожного виду документа, пов'язаного з продукцією, метою якого є розкриття інформації про продукт або послугу. Технічна документація поділяється на внутрішню і зовнішню. Під внутрішньою документацією розуміють технічні креслення, переліки деталей, переліки робіт, робочі інструкції та ін. Вона є основоположною при розробці, створенні та обслуговуванні продукції. Зовнішня документація, що включає технічні паспорти, каталоги запасних частин і керівництва, адресована існуючим клієнтам і частково використовується для їх придбання. Зовнішня документація також

підтверджує специфікації продукції для державних органів. Різні види вимагають наявності технічної документації для виконання певних вимог, наприклад:

– аудиторія повинна бути відома і відповідним чином адресована;

– мовний стиль повинен бути спрямований на розуміння описуваного питання;

– документація повинна бути повною і структурованою відповідно до потреб користувачів;

– необхідно дотримуватися законів і стандартів;

– необхідно знайти адекватний баланс між картинками і текстом;

– документ повинен бути привабливим і як можна більш коротким.

Всі технічні документи підкреслюють зрозумілість як ключовий аспект своєї структури. Головна мета технічної документації полягає в тому, щоб кінцеві користувачі, а також співробітники різних відділів компанії могли зрозуміти її без додаткових досліджень. Це гарантує, що тексти в основному написані ясно, просто і лаконічно і не містять занадто багато скорочень або внутрішньої корпоративної термінології. Крім того, технічна документація зазвичай багаторазово перевіряється перед публікацією або для забезпечення її якості за допомогою коректорів, або для забезпечення її зрозумілості за допомогою аналізу аудиторії. Проте, аналіз технічної документації пов'язаний з багатьма труднощами для систем машинного перекладу, оскільки використовувана мова є суто технічною і не всі скорочення можна уникнути.

Щоб забезпечити відповідність високим вимогам якості і бути юридично захищеним, необхідно дотримуватися специфічних стандартів на продукцію, щоб досягти відповідності з відповідними законами і уникнути можливих компенсаційних претензій. В рамках Європейського співтовариства численні обов'язкові директиви ухвалені в національному законодавстві і часто доповнюються основними стандартами галузі. На закінчення слід зазначити, що великі вимоги, що пред'являються до технічної документації, повинні забезпечити

високу якість текстових частин в складному середовищі і, отже, створити міцну, але складну основу для систем машинного перекладу.

Висновки до розділу

Метою даної роботи була оцінка якості технічних документів та їх перекладів за допомогою методів машинного навчання з акцентом на виявлення відмінностей між автоматичними перекладами і професійними перекладами , виконаними людьми. Розглядаються два дослідних питанняє.

1. Як можна оцінити якість перекладу технічної документації, якщо є оригінал документа?

2. Як можна оцінити якість перекладу технічної документації, якщо оригінал документа недоступний?

Щоб відповісти на ці питання, вони були розбиті на наступні практичні кроки.

1. Надати алгоритм машинного навчання з оптимальною якістю передбачення для ідентифікації професійних і автоматизованих перекладів технічних документів з доступом і без доступу до оригіналу документа.

2. Ввести належну основу для ранжирування якості технічних документів і фрагментів технічних документів незалежно від типу їх перекладу.

Такий поділ було вибрано тому, що перші кроки, включаючи створення і попередню обробку даних, досить схожі в двох дослідницьких питаннях. Тому ефективніше відразу побудувати систему для алгоритму машинного навчання, який вирішує поставлену задачу отримання даних зі знанням і без знання вихідного документа. Основна відмінність між цими двома варіантами полягає у виборі атрибутів, причому для установки зі знанням допускається значно більше змінних. Створення нової структури для ранжирування якості технічних документів буде будуватися на основі результатів першого робочого кроку. Таким чином, процес

виявлення знань зосереджений на першому робочому кроці, а другий вирішується з використанням отриманих результатів.

Важливим обмеженням, є те, що дана робота зосереджена на оцінці синтаксису і не враховує семантичні частини тексту, оскільки це виходить за рамки мети даної роботи.

Розділ 3
Розробка системи класифікації текстових документів за формальними претендентами

У цьому розділі розглянемо кроки, необхідні для створення системи оцінки машинного перекладу з наголосом на технічну документацію. Для цього отримаємо відповіді на питання.

- Як можна відкрити нові знання?
- Які цілі і обмеження впливають на обраний спосіб?
- Який метод дає можливість відповісти на питання дослідження і як він складений?
- Які атрибути даних необхідні для застосування методу?
- Як можна застосувати відкриті знання в подальшому?

Для того щоб відповісти на ці питання, структура цього розділу відображає адаптацію раніше представленого процесу виявлення знань. Виконаний процес проілюстрований на рисунку 3.1.

З огляду на ітеративну природу процесів виявлення знань, робота розділена на два основних цикли:

1. Перший цикл включає в себе ЗБД-кроки 2 і 3 (див. розділ 2). На відміну від звичайного знання процеси виявлення, набір даних створювався і відповідні атрибути вибиралися вручну. Це призвело до високої залежності між етапом попередньої обробки і етапом збору даних, оскільки оптимізація набору даних шляхом видалення помилкових записів даних або поліпшення загальної якості елементів даних може змінити придатність вибраних атрибутів або збільшити потрібний обсяг даних. Крім того, при конфігурації набору даних приймається рішення про розміри кожного запису даних. Після вибору відповідних атрибутів може знадобитися повторне коригування довжини для більш точної відповідності атрибутам. Це створює необхідність ітеративного процесу між вибором або створенням відповідного набору даних, що містить перекладені фрагменти тексту, і

вибором атрибутів.

2. Другий цикл включає в себе ЗБД-кроки 2-8, що дозволяє перенастроювати набір даних, попередня обробка, вибір атрибутів або алгоритм машинного навчання. Цей цикл ще більш важливий для роботи, оскільки основною метою є оптимізація якості прогнозування.

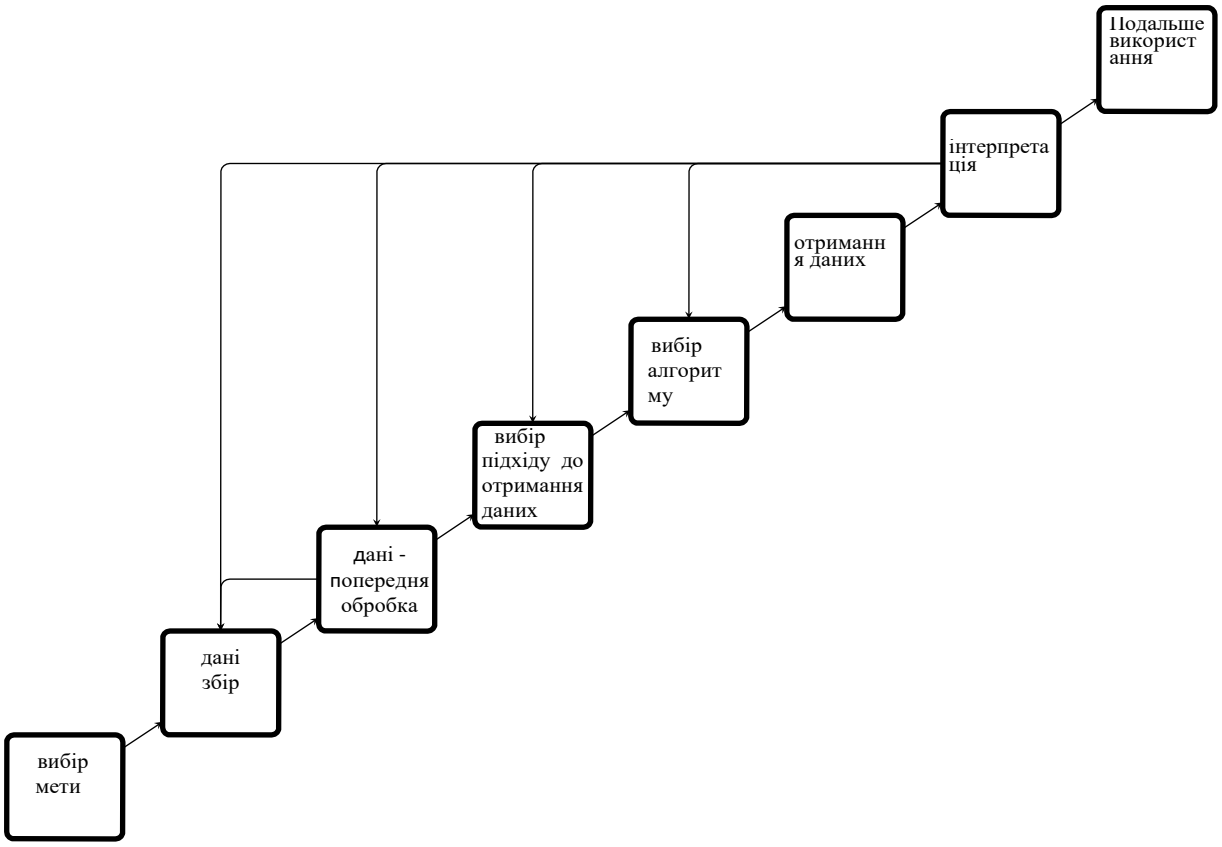


Рисунок 3.1 – Адаптований процес виявлення знань, що містить кроки від збору даних до інтерпретації результатів і їх використання для вирішення подальших завдань

Для цього використовуємо представлений процес ітеративно, тестуючи кілька алгоритмів для порівняння їх один з одним. Крім того, результати роботи алгоритмів можуть вказувати на невдало обрану базу даних і, отже, створювати необхідність коригування на етапах 2 і 3.

3.1 Отримання тексту і поділ пропозицій

Для того щоб оцінити машинний переклад і почати новий процес виявлення знань, необхідно зібрати достатню кількість вихідних даних і задати базову структуру даних.

Класичні задачі отримання даних вимагають наступних типів даних: навчальний набір даних, який використовується для навчання алгоритму на конкретних даних завдання, і тестовий набір даних для оцінки якості алгоритму. Дана задача спрямована на бінарну класифікацію технічної документації, класами якої є автоматичний і професійний переклад. Тому на етапі збору даних необхідно було створити як мінімум професійний, людський переклад та автоматизований переклад для кожного тексту бази даних. У зв'язку з тим, що багато хто з представлених метрик вимагають еталонного перекладу для оцінки фрагмента тексту-кандидата, необхідно надати додатковий переклад, який можна розглядати як еталон для порівняння з кандидатами. На закінчення слід зазначити, що остаточний набір даних перекладів містить наступні частини для кожного запису:

– ригінальний текст;

– професійний або автоматизований переклад як кандидат на оцінку;

– один або кілька еталонних перекладів одного і того ж оригінального тексту, що використовуються для підрахунку балів. Алгоритми оцінки класичних машинних перекладів часто вимагають високої якості довідникових текстів. Потім ці посилання використовуються алгоритмами, такими як АДО або Meteor, для розрахунку балів якості.

Отже, документи для збору вже повинні бути доступні в оригінальній версії і однієї або декількох відповідних професійно переведених версіях, оскільки створення додаткових професійних перекладів текстів виходить за рамки мети даної роботи.

Перші спроби збору даних показали відсутність вільно доступних даних для проведення оцінки на рівні документів. Зважаючи на наявність документів і великої кількості даних, необхідних для ефективного пошуку даних, документи були розбиті на речення, і набір даних був побудований на основі цієї меншою логічної одиниці - рівня речення. Це розширило мету даної роботи - надати системі умовної класифікації, здатної бінарно класифікувати речення по двом вищезгаданим міткам. Шляхом рекомбінації речень в вихідні документи, оцінка якості документа все ще можлива. Таке розбиття значно збільшує набір даних, але створює додаткову проблему збору перекладів документів за перекладом речень.

Після ретельного дослідження з'ясувалося, що відповідну кількість технічних документів, що знаходяться у вільному доступі, включаючи професійний переклад, є тільки в форматі PDF. Оскільки PDF є форматом документів, орієнтованим на візуалізацію, витяг з нього окремих речень не є простим завданням. Далі докладно описані кроки, виконані для отримання і порівняння речень.

Спочатку наявні PDF-файли перетворюються в звичайні текстові файли. Через особливості формату PDF отриманий текстовий файл містить безліч помилок і не є логічно впорядкованим, наприклад, виноски вбудовуються в речення, а речення або слова розбиваються і розділяються. Корисною характеристикою використаних документів були легко помітні абзаци.

Тому звичайні текстові файли були розділені на відповідні абзаци, де абзаци визначаються як набір декількох послідовних рядків, що містять текст, без проміжних порожніх рядків. Для того щоб знайти і видалити абзаци, які містять реальні речення, дослідження показало, що існуючі абзаци можна відфільтрувати по трьом вимогам. Вони повинні складатися як мінімум з 15 символів, трьох прогалин і одного розділового знака. В іншому випадку вони, швидше за все, містили не справжній текст, а текст, що не придатний для використання, наприклад, виноски, заголовки і підписи до малюнків.

Переконавшись, що залишилися абзаци, які не містять нічого, крім корисного контенту, вони були розділені на окремі речення за допомогою алгоритмів

тегування частин мови. Визначення тегів частин мови в реченні дозволяє розбивати речення незалежно від пунктуації, що дуже корисно в контексті технічної документації. Це пов'язано з їх технічною природою, в результаті чого в документах виникає безліч технічних розділових знаків, що робить неефективним розбиття речень на знаки пунктуації.

Після цього витягнуті речення аналізувалися далі, щоб видалити кінцеві фрагменти і переконатися, що витягнуті рядки дійсно утворюють повні речення. Таким чином, речення були відфільтровані на наявність більше двох прогалин, на закінчення з пунктуацією і містили не менше семи символів. Додаткові вимоги, такі як наявність дієслів речень, виявилися неефективними, оскільки опрацьовані документи містили численні списки і торгові марки, які заважали цим правилам.

Щоб завершити збір вихідних даних, витягнуті речення необхідно було пов'язати з відповідним перекладом. Однак не вдалося встановити процес вилучення і розділення речень, який гарантував би створення однакової кількості витягнутих речень. Тому пропозиції були перекладені системами автоматичного перекладу, які, в свою чергу, були використані для створення процесу, який визначає, чи можуть два перекладені речення бути результатом одного і того ж оригінального документа. Це дозволило впорядкувати набір даних, співставивши відповідні речення один з одним і видаливши речення, які не мають збігів з жодною зі сторін.

Нарешті, для кожного речення було створено три незалежних машинних переклади, щоб створити достатню кількість автоматичних перекладів і вирішити проблему відсутності професійного еталонного перекладу. Оскільки створити додатковий професійний переклад для даного набору даних не представлялося можливим, використовували два підходи для вирішення цієї проблеми.

По-перше, як еталон був використаний один з наборів автоматизованого перекладу. Більшість метрик оцінки перекладу припускають наявність високоякісного еталона для порівняння речень. У наведеному прикладі показник АДО розраховує значення подібності, причому більш високе значення означає вищу схожість з даними еталона. Це означає, що якщо еталон не відрізняється високою

якістю, то схожість речення низької якості з еталоном може бути вище, ніж схожість речення високої якості. Оскільки перше питання дослідження спрямоване на вирішення проблеми бінарної класифікації, а не на оцінку якості заданих текстів, посилання не обов'язково повинне бути високої якості. Очікуваний результат полягає в тому, що при використанні автоматизованого перекладу як еталону, автоматизовані тексти-кандидати отримають вищі бали подібності, ніж професійні.

Другий підхід заснований на висновках. Поєднання кількох псевдоспосилань може привести до високоякісного посилання, аналогічному перекладу, створеному людиною. Таким чином, псевдореференс - це посилання, яке не було створене людиною, в даному випадку програмою машинного перекладу. Тому для формування еталонного перекладу були використані два автоматичних переклади в комбінації. В цьому випадку очікуваними результатами були вищі оцінки подібності для професійних перекладів, ніж для автоматичних, оскільки очікувалося, що комбінація створить еталон вищої якості.

Крім того, для створення додаткових атрибутів, які не потребують знання оригінального документа, був використаний метод переведення туди і назад для створення другої версії документа.

Пропозиція-кандидат, яка буде називатися референцією туди і назад. Це було зроблено шляхом перекладу даного речення-кандидата на іноземну мову і назад на мову оригіналу. При цьому речення було явно змінено, так як переклад в обидві сторони зазвичай призводить до отримання помилкових даних. Але це дозволило створити еталонний фрагмент тексту, який можна було використовувати в якості порівняння для перекладу-кандидата без знання вихідного документа. Таким чином, алгоритми, які зазвичай вимагають еталонного перекладу, могли бути розраховані і для дослідницького питання, що не допускає використання оригінального документа. Для створення достатньої кількості автоматизованих перекладів були використані наступні системи машинного перекладу:

- SDL freetranslations.com;
- GoogleTranslate;

– Bing Microsoft.

Щоб зменшити можливі залежності від обраних автоматичних перекладачів, на подальших етапах використовуються дев'ять різних видів комбінацій даних, де кожен машинний переклад використовується один раз в якості кандидата, а два інших використовуються окремо і в комбінації і в якості еталонного перекладу. Однак через необхідність використання еталонного перекладу в обидві сторони третя частина наборів даних, що містять систему Freetranslation як кандидата, була видалена на ранніх етапах процесу оптимізації. На наступному рисунку показаний набір даних, який в подальшому буде називатися обробленим набором перекладів:

Таблиця 3.1 – Загальна структура набору даних

кандидат	посилання 1	посилання 2
Проф. транс. 1	Авто. Trans. 1 Google Translate	Авто. Trans. 1 вільний переклад
Авто. Транс. 1 Bing Translator	Авто. Trans. 1 Google Translate	Авто. Trans. 1 вільний переклад
Проф. транс. 2	Авто. Trans. 2 Google Translate	Авто. Trans. 2 вільний переклад
Авто. Транс. 2 Bing Translator	Авто. Trans. 2 Google Translate	Авто. Trans. 2 вільний переклад

3.2 Система вибору атрибутів

Вибір атрибутів є важливою частиною виявлення знань. Метою атрибута є отримання знань в тій чи іншій формі щодо кожного запису даних. Як правило, атрибут - це змінна, яка має значення для кожного запису набору даних і може бути обчислена або доступна в тій чи іншій формі. Крім того, важливим фактором є обраний тип атрибута, оскільки не кожен алгоритм може працювати з будь-яким типом атрибута. Два основних типи - це номінальні і числові атрибути. Номінальні атрибути складаються з фіксованого набору зазвичай не сортованих значень, в той

час як числові мають безперервну і нескінченну природу. Для отримання додаткових знань про використовувані метрики, таких як показник Meteor, були додані проміжні етапи процесу розрахунку. Далі будуть коротко описані вибрані атрибути:

– модифікована оцінка уніграмм АДО.;

Алгоритм двомовної оцінки АДО є швидким і недорогим методом оцінки якості машинного перекладу повністю автоматизованим способом шляхом обчислення значення від 0 до 1 в залежності від "близькості" еталона і кандидата. В уніграммній оцінці АДО використовуються 1-грами , тобто не враховуються послідовності слів, а тільки окремі слова для розрахунку модифікованої оцінки АДО.

– модифікована оцінка АДО біграмм і триграм;

Для того щоб дослідити різницю між АДО-оцінками одного слова і АДО-оцінками послідовності слів для даного набору даних, були розраховані АДО-оцінки для 2-грам і 3-грам і додані в якості атрибутів.

– Meteor бал;

Meteor - це версія оцінки АДО, яка була оптимізована для роботи на рівні пропозицій.

– повнота Meteor;

Показник повнота Meteor описує кількість схожих уніграмм в перекладі-кандидата і еталонному перекладі по відношенню до кількості уніграмм в фрагменті-кандидаті.

– Meteor точність;

Показник точності Meteor характеризує кількість однакових уніграмм в перекладі-кандидата і еталонному перекладі по відношенню до кількості уніграмм в еталонному фрагменті.

– Meteor частки;

Оцінка Meteor Chunk підраховує мінімально необхідну кількість фрагментів для кожного пропозиції.

– штраф за фрагменти Meteor;

Штраф за фрагментацію розраховується таким чином, щоб надати більшої ваги збігів з довшими n-грамами. Це означає, що речення з великою кількістю довших n-грам в перекладі кандидата і еталона вимагає меншої кількості фрагментів, що в свою чергу призводить до зниження штрафного бала.

– Meteor F1;

Показник F1 розраховує середнє гармонійне між точністю і повнотою.

– середній Meteor;

Показник Meteor Mean розраховує середнє гармонійне між точністю і повнотам, при цьому Повнота має в дев'ять разів більшу вагу, ніж точність.

– Meteor збігів;

Показник загальної кількості збігів підраховує загальну кількість збігів, знайдених між еталонним і перекладом-кандидатом на основі уніграмм.

– контрольна довжина;

Оцінка довжини посилання порівнює довжині перекладу кандидата з довжиною перекладу посилання, генеруючи різницю між двома фрагментами тексту й зберігаючи її в якості значення атрибута. Це призводить до від'ємних показників для більш коротких кандидатів і позитивним - для більш довгих.

– оцінка частин мови;

Оцінка частин мови - це логічне значення для опису того, чи відповідає переклад-кандидат заданому мінімальному шаблону необхідних тегів частин мови для формування граматично правильного речення.

– легкість читання;

Алгоритм зручності читання Флеша розраховує бал, який вимірює читаємість речень і документів.

– оцінка редагування перекладу;

Коефіцієнт редагування перекладу вимірює кількість правок, необхідних для перетворення тексту-кандидата в еталонний переклад, щодо довжини тексту.

– редагування перекладу правок;

Проміжний етап підрахунку загальної кількості правок був реалізований для отримання інформації про кількість правок без урахування довжини посилання. Завдяки тому, що не використовується відносна оцінка, враховується довжина перекладу-кандидата і еталона, що дає додаткові знання для алгоритму машинного навчання.

– використані посилання;

В деяких частинах експерименту для створення більш стандартизованої еталонної версії використовувалося кілька еталонних перекладів. Більшість метрик, таких як Meteor і АДО, здатні впоратися з декількома еталонними перекладами, розраховуючи оцінки для всіх еталонів і вибираючи кращий. відстежували, який переклад посилання використовувався частіше при розрахунку всіх метрик, щоб отримати додаткові знання, які можуть полегшити процес класифікації для даного алгоритму машинного навчання.

– підрахунок помилок;

Для аналізу фрагментів тексту кандидата на предмет їх словесної і граматичної правильності була використана система перевірки стилістичних слів і граматики (language tool). Програма підраховує всі знайдені помилки і розділяє їх на різні категорії. На додаток до загальної кількості помилок були додані окремі категорії. Однією з основних проблем, що виникли на цьому етапі, був пошук потрібної кількості атрибутів, відповідних вимозі відсутності знань про вихідний документ. Видалення вихідного документа як ресурсу перетворює завдання з завдання машинного перекладу в задачу аналізу якості тексту, оскільки алгоритмам більше не доступна інформація про те, що даний текст є перекладом.

Таким чином, область машинного перекладу не є безпосередньо застосовною для вирішення цієї частини дослідницького питання. Основними властивостями для хорошої якості тексту є граматична правильність, правильність слів, семантична правильність і стиль. Як уже згадувалося семантична правильність речення виходить за рамки завдань даної роботи і тому не може бути прийнята до уваги. Це залишає частину дослідницького питання без знання оригінального документа з

трьома типами проблем, які необхідно вирішити:

1. Граматична правильність речення або тексту може бути перевірена або оцінена. Зазвичай це робиться за допомогою програм, заснованих на правилах, які перевіряють речення на порушення цих правил і повідомляють про відповідні помилки.

2. Правильність слова оцінити складніше, ніж граматичну правильність для даного завдання. Фокус на технічну документацію обумовлює деякі властивості документів, які ускладнюють ці оцінки.

По-перше, технічна природа, що призводить до появи безлічі невідомих слів для звичайних словників, а по-друге, технічна документація часто присвячена конкретному продукту компанії і тому містить безліч власних іменників, які навіть не можуть бути виявлені словниками , спеціалізованими для технічної документації.

3. Стиль речення в цілому не застосовується для виявлення помилкових речень, однак метрики, такі як алгоритм читабельності, можуть допомогти в генеруванні інформації про якість документа.

Як описано вище, можливості оцінки якості документа без знання вихідного тексту обмежені. На відміну від цього, продуктивність і якість передбачення алгоритму машинного навчання сильно залежать від доступних атрибутів для навчання і тестування. Це є проблемою для вирішення даної частини питання дослідження, яка може поставити під загрозу його успіх. Для вирішення цієї проблеми було створено еталон «туди-назад». Це дозволило створити додаткові атрибути, згадані вище, без використання вихідного документа,

На етапі попередньої обробки атрибутів основним завданням для даної задачі є виявлення викидів серед значень атрибутів. Значення вважається викинутим, якщо воно відхиляється на значну величину від середнього значення найближчих значень. Для даної задачі викиди були обчислені за допомогою коефіцієнтів викидів класу, які розраховуються шляхом ранжирування кожного примірника набору даних

Основною перевагою нормалізації є порівнянність, що в основному стосується метрик перекладу таких як АДО, Meteor і КРП, які нормалізовані за замовчуванням. Однак для виявлення викидів важливо мати сумісність усіх існуючих атрибутів, тому для даного набору даних виконується перетворення діапазону, відображаючи дані в значення від 0 до 1.

Для того щоб визначити, чи потрібно і які ітераційні цикли робити після виконання етапу отримання даних, необхідні певні метрики для оцінки результатів різних алгоритмів. Першою і найбільш значущою метрикою є показник неправильної класифікації. Оскільки для даного набору даних розподіл міток є абсолютно рівномірним, низький показник помилкової класифікації або, відповідно, високий показник точності можна розглядати як результат роботи алгоритму дуже високої якості. Щоб підвищити достовірність точності алгоритму, результати були співставлені з результатами фіктивного класифікатора, як відповідного зразка. Фіктивний класифікатор завжди пророкує найбільш поширений клас для всього набору даних, в результаті чого точність класифікації становить 50%. Обчисливши різницю між результатами алгоритмів і фіктивного класифікатора, був розрахований приріст точності, щоб отримати більш чітке уявлення про додаткові знання, отримані при використанні алгоритмів. Хоча розподіл класів в даному наборі даних однаковий, все ж представляє великий інтерес визначити, чи є алгоритм значно краще в виявленні автоперекладів або значно краще у виявленні професійних перекладів. Це можна зробити, відстежуючи досягнуту точність і повноту кожного алгоритму. Тільки обчислення точності недостатньо, оскільки, хоча точність дає кількість правильно класифікованих примірників для одного примірника по відношенню до всіх екземплярів, класифікованих з цією міткою, висока точність може бути також досягнута за рахунок класифікації відповідної мітки тільки в дуже специфічних сценаріях і, таким чином, буде упущення багатьох випадків, які дійсно відносяться до цієї мітки. Для боротьби з цим можна використовувати показник повнота в поєднанні з показником точність. Оскільки показник повнота вимірює кількість правильно класифікованих примірників по

відношенню до всіх екземплярів, які дійсно належать до даного класу, описана вище стратегія отримає значно менший показник повноти. Для того щоб об'єднати обидва значення, використовується F-середнє, обчислюючи середнє гармонійне між двома оцінками, в результаті чого виходить значення для оцінки здатності алгоритму класифікувати певний клас. Щоб врахувати обидва класи, F-середня була розрахована для обох міток.

3.3 Розробка системи класифікації на основі документів

Останнім етапом будь-якого процесу виявлення знань є застосування отриманих результатів для подальшого використання. У запропонованій задачі це особливо важливо, тому що основним напрямком даної роботи була класифікація перекладів на рівні документів і, крім того, оцінка перекладів. В даному розділі пояснюється, як результати ЗБД-процесу використовуються для виконання класифікації на рівні документів і як створити структуру для оцінки переведених речень і документів.

Для того щоб передбачити оригінальні документи, окремі речення повинні бути перекомбіновані в відповідні технічні документи. В результаті для кожного технічного документа створюється професійна і автоматизована версія. Щоб класифікувати документ з певною міткою, всі речення, які стосуються цього конкретного документу, використовувалися в якості резервних, а решту документів об'єднувалися в навчальний набір даних для алгоритму. (Наприклад, професійно перекладений документ під назвою "установка обладнання" використовується в якості резервного набору, а решту документів об'єднуються для створення навчального набору для алгоритму). Потім алгоритм класифікує кожне речення утримуваного набору з певною міткою, а документ класифікується з найбільш поширеною міткою серед його речень. Цей процес повторюється для кожної версії кожного технічного документа, а коефіцієнт помилкової класифікації

розраховується шляхом віднесення кількості правильно класифікованих документів до всіх документів.

У зв'язку з невеликою кількістю доступних документів було здійснено другий крок для підтвердження достовірності результатів. Створили додаткову базу даних технічних документів, випадковим чином комбінуючи речення в фіктивні документи. Рисунок 3.2 ілюструє процес створення додаткових фіктивних документів.

Крім того, цей підхід дозволив вивчити необхідний розмір технічних документів. Створюючи документи суттєво різних розмірів, можна було визначити мінімальну довжину технічних документів, щоб з високою ймовірністю правильно їх класифікувати. Крім того, цей підхід дозволив визначити необхідний рівень поширення речень в документі залежно від його розміру, щоб класифікувати їх з певною міткою.

Описані метрики були інтегровані з використанням пакетів з відкритим вихідним кодом. Більшість метрик були адаптовані відповідно до вимог конкретного завдання машинного навчання і розраховані для формування остаточної бази даних, використовуваної для навчання і перевірки алгоритму машинного навчання.

Перед оптимізацією кожного алгоритму набір даних був попередньо очищений, щоб перетворити його в оптимальні умови для навчання алгоритму. Ці кроки, включаючи процес оптимізації, були виконані за допомогою програмного забезпечення RapidMiner і детально показані на рисунку 3.3.

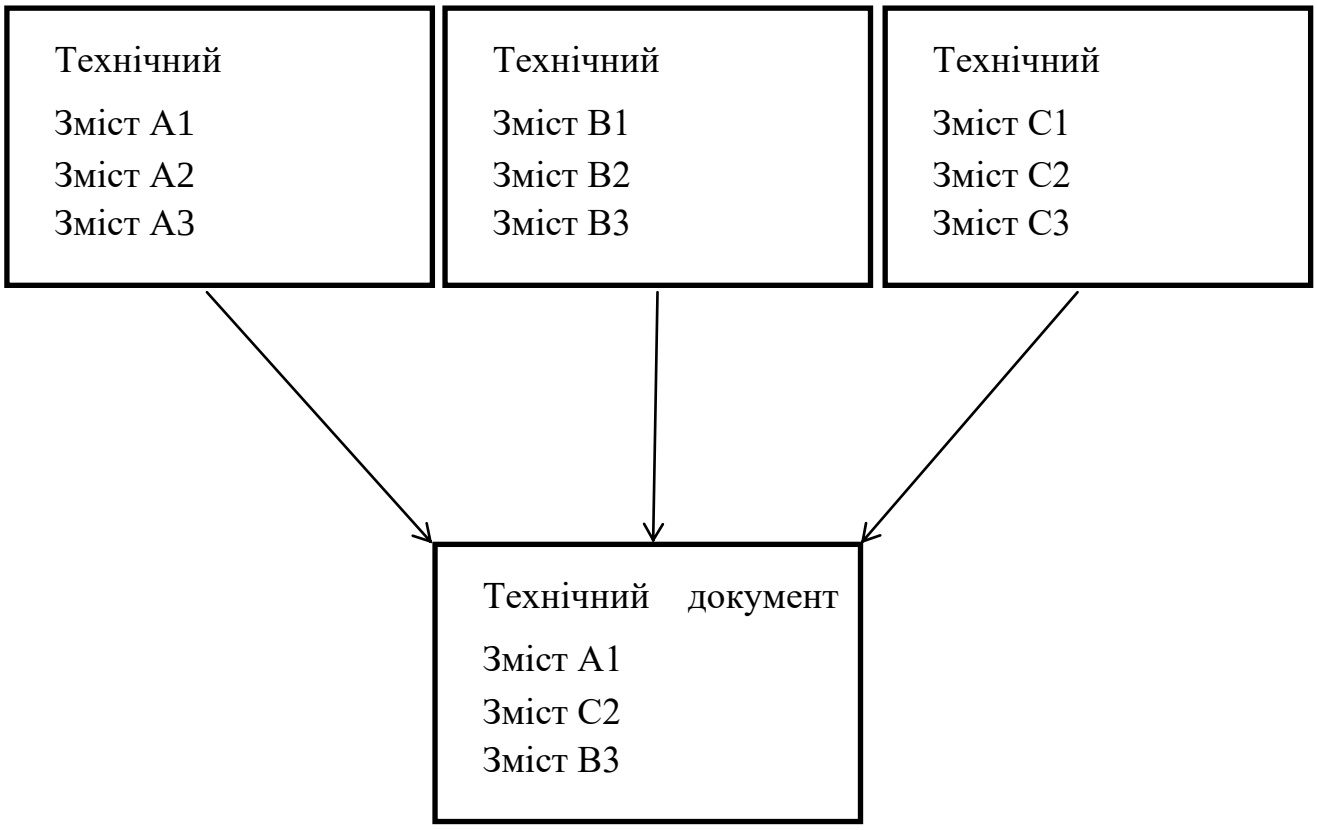


Рисунок 3.2 – Створення технічного документа

Після зчитування вхідного набору даних і установки типів атрибутів виконується перший етап попередньої обробки - видалення дублікатів. Цей крок видаляє всі пов'язані записи в наборі даних, крім однієї, що важливо для обмеження впливу кількох однакових записів даних. Далі дані нормалізуються за допомогою перетворення діапазону. При цьому всі атрибути приводяться до одного і того ж діапазону значень, що дозволяє легше порівнювати їх між собою і робить можливим наступний крок - виявлення викидів.

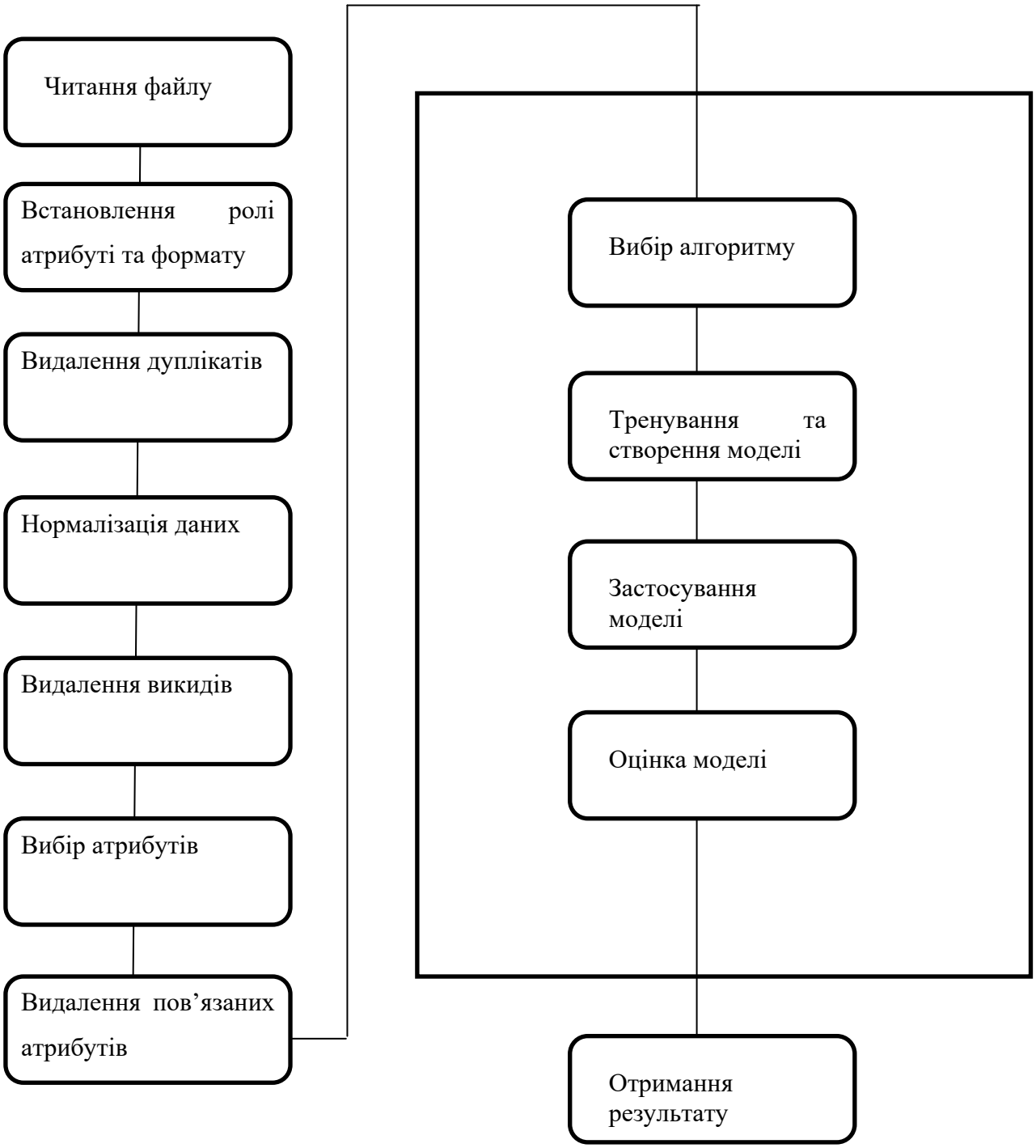


Рисунок 3.3 – Попередня обробка набору даних і оптимізація алгоритму машинного навчання

Викиди визначаються за допомогою коефіцієнтів викидів класу шляхом порівняння записів даних один з одним за допомогою метрики евклідова відстань. 5% найбільш відмінних записів даних визначаються як відхилені значення і не розглядаються для завдання машинного навчання. Крім того, в залежності від

питання дослідження вибираються атрибути, допустимі для даного завдання, що включає видалення 12 атрибутів для другого питання дослідження. Останнім кроком попередньої обробки для подальшого скорочення часу обчислень і усунення надмірності атрибутів є виявлення висококорельованих атрибутів і їх видалення, якщо кореляція перевищує 90%.

Висновки до розділу

Практичним робочим кроком для відповіді на поставлені завдання стала структура речень для оцінки технічної документації без урахування типу її перекладу. Використовуючи результати процесу виявлення знань, можна було створити класи, що дозволяють поділяти речення і, отже, документи в залежності від якості їх перекладу. Для оцінки використовувалися три властивості процесу класифікації: передбачена мітка для двох найбільш ефективних алгоритмів і кількість помилок, що відслідковуються для речення. Щоб отримати більш адекватну репрезентацію передбаченої мітки і відслідковування помилок, кількість помилок було розділена на більш тонкі атрибути, в той час як передбачення класів для двох алгоритмів залишилися незмінними. Отримані категорії були проаналізовані на предмет їх впливу на зниження якості речення. Кожна категорія була зважена в залежності від її серйозності, надаючи помилкам з високим впливом додаткової ваги, а з низьким впливом - менший або навіть не надаючи ніякої ваги, якщо помилка не мала значення для якості пропозиції.

Передбачення того, що речення було професійно переведено, розглядалося як високий показник якості речення, оскільки в загальному випадку людський переклад являє собою більш високу якість, ніж при автоматичному перекладі.

Таким чином, трьома результуючими властивостями для оцінки якості речення були два прогнози, зроблені за допомогою інтелектуального аналізу даних, і зважений бал його помилок. Реченню присвоювався клас якості залежно від значень трьох згаданих атрибутів.

Розділ 4
Дослідження ефективності методів класифікації текстових пропозицій за формальними методами

4.1 Чисельні результати порівняльної класифікації

В результаті вилучення речень з технічних документів було отримано 30000 рядків, що містять фрагменти тексту. Оскільки витяг речень не є простим завданням, 8000 з цих рядків не були витягнуті правильно, що призвело до появи помилкових фрагментів тексту, що не утворюють правильних речень. Таким чином, кожен кінцевий набір даних, який використовувався для навчання і тестування алгоритмів машинного навчання, складався з 22327 речень для кожної системи перекладу.

Таблиця 4.1 – Статистика використаних наборів даних

Кількість систем автоматизованого перекладу	3
Кількість технічних документів	15
Початкова кількість речень у всіх документах	30
Кількість речень, використаних кожною системою перекладу для наборів даних	327
Приблизна кількість речень в наборах даних	34, 730
Загальна кількість наборів даних	7
Кількість атрибутів, необхідних для еталонного перекладу	14
Кількість атрибутів, які не потребують перекладу посилань	19
Загальна кількість атрибутів	42
Максимальна кількість доступних значень атрибутів	1 456 210
Кількість створених фіктивних документів	190

Щоб забезпечити рівномірний розподіл між мітками "професійний переклад" і "автоматичний переклад", кожен набір даних був складений з двох наборів речень,

один з яких переводився професійно, а інший – автоматично. В результаті кожен набір даних містив в цілому 42458 речень.

Для подальшого вивчення валідності алгоритму на різних розмірах технічної документації, виготовлені документи варіювалися в розмірах від 5 до 3500 речень в документації. Важливо відзначити, що створення документів більшого розміру було більш складним завданням через обмежену кількість речень і необхідного обсягу навчальних даних для створення значимої моделі класифікації. Крім того, інформація, отримана з документів меншого розміру, є більш цінною, так як необхідну кількість речень для правильної класифікації документа, як очікується, буде варіюватися від 45 до 200 речень.

У таблиці 4.2 представлені дев'ять використаних наборів даних, створених шляхом використання трьох систем машинного перекладу в якості кандидатів і референсів відповідно і створення наскрізного перекладу для шести з дев'яти наборів даних.

Щоб оцінити вплив вищезгаданих кроків, процес оптимізації був сформований для різних налаштувань шляхом включення і виключення певних кроків, наприклад, Видалення дублікатів = true, нормалізація даних = false, видалення викидів = true, видалення корельованих атрибутів = false. Ітерація оптимізації складається з установки параметрів алгоритму, побудови моделі, застосування збереженого набру до побудованої моделі та оцінки результатів. Набір для порівняння був обраний випадковим чином як 30% від набору даних. Цей процес був інтегрований в автоматизовану установку для виконання декількох процесів оптимізації на основі еволюції. Перший крок для кожного алгоритму складався з невеликого кроку оптимізації для вивчення впливу різних налаштувань підготовки даних і загальної придатності алгоритму для даної задачі. Експерименти проводилися на всіх восьми наборах даних, і виявилося, що підготовка даних була корисною для кожного налаштування, тому на наступних етапах оптимізації застосовувалися нормалізація, видалення дублікатів, виявлення викидів і видалення корельованих атрибутів.

Таблиця 4.2 – Використовувані комбінації референтів і кандидатів

Кандидат	Посилання1	Посилання2	Переклад в обидва кінці
Google Translate	Bing	- -	Google ПТН через Freetranslation
Google Translate	вільний переклад	- -	Google ПТН через Freetranslation
Google Translate	Bing	вільний переклад	Google ПТН через Freetranslation
перекладач Bing	Google	- -	Bing ПТН через Freetranslation
перекладач Bing	вільний переклад	- -	Bing ПТН через Freetranslation
перекладач Bing	Google	вільний переклад	Bing ПТН через Freetranslation
вільний переклад	Google	- -	- -
вільний переклад	Bing	- -	- -
вільний переклад	Google	Bing	- -

Крім того, з подальшим процесом оптимізації були виключені допоміжні SVM і класифікатор Naive Bayes, оскільки SVM показав значно гірші результати, ніж інші алгоритми, а Naive Bayes не мав можливості подальшої оптимізації через відсутність адаптування параметрів. Наступний крок полягав в більш глибокій оптимізації шляхом створення більшої кількості моделей для кожного алгоритму і тестування більшого діапазону налаштувань параметрів. У таблиці 4.3 наведено огляд згенерованих моделей для обох питань дослідження і відповідних оптимізацій.

Таблиця 4.3 – Огляд створених моделей і оптимізацій

	Дослідницький	Дослідницький	**всього**
Побудовані дерева рішень	320 000	200 000	680 000
Побудовані нейронні мережі	5 000	1 000	7 000
Побудова k-найближчих сусідів	3 100	1 500	54 000
Побудовані опорні SVM	2 200	1 000	2 500
Загальна кількість оптимізацій	323 000	204 500	694 500

Для короткого огляду в таблиці 4.4 приведена сукупність наступних більш докладних результатів і показана загальна продуктивність використаних алгоритмів.

Таблиця 4.4 – Середні значення та стандартні відхилення протестованих алгоритмів

алгоритм	точність	Стандартне відхилення
дерево рішень	67.34%	0.013
Штучна нейронна мережа	**72.45%**	0.015
k-Nearest Neighbor	68.45%	**0.010**
наївний Байес	67.67%	0.020
SVM з підтримкою	63.76%	0.018

У таблиці 4.5 показані результати з найвищою точністю передбачення речень за допомогою Дерев рішень, включаючи відповідні F1. Для кожної комбінації кандидата і посилання, наведеної в таблиці, було побудовано і оцінено 50 000 дерев рішень. Дані, що використовуються включали всі доступні атрибути в нормалізованому вигляді, зменшені на 5% шляхом виявлення викидів і подальшого видалення дублікатів.

У наступній таблиці представлені найвищі результати для штучних нейронних мереж. Кожна комбінація кандидат-референт була оптимізована 100 разів. Підготовка даних аналогічна попередній оптимізації Дерева рішень.

Таблиця 4.5 – Огляд кращих результатів використання дерева рішень для відповідних комбінацій кандидата і еталона

Дерево рішень					
кандидат	посилання 1	посилання 2	точність	F1-автоматизований	F1-професійний
Google	Bing	- -	68.19%	0.705	0.743
Google	вільний переклад	- -	64.93%	0.758	0.683
Google	Bing	вільний переклад	**73.43%**	0.738	0.684
Bing	Google	- -	72.19%	0.782	0.662
Bing	вільний переклад	- -	67.76%	0.737	0.658
Bing	Google	вільний переклад	68.54%	0.783	0.683
вільний переклад	Bing	- -	64.21%	0.762	0.651
вільний переклад	Google	- -	63.63%	0.652	0.628
вільний переклад	Bing	Google	66.73%	0.689	0.649

Таблиця 4.6 – Огляд кращих результатів використання штучної нейронної мережі для відповідних комбінацій кандидат-референт

Штучна нейронна мережа					
кандидат	посилання 1	посилання 2	точність	F1-автоматизований	F1-професійний
Google	Bing	- -	70.43%	0.729	0.681
Google	вільний переклад	- -	66.65%	0.727	0.681
Google	Bing	вільний переклад	71.34%	0.756	0.671
Bing	Google	- -	**73.25%**	0.721	0.724
Bing	вільний	- -	69.82%	0.681	0.701
Bing	Google	вільний переклад	71.84%	0.714	0.723

Кожна комбінація містить всі доступні атрибути. Атрибути нормалізуються, і набір даних скорочується до 5% викидів. Дублікати не враховуються.

Результати для використаного алгоритму k-найближчих сусідів показані в таблиці 4.7. Для кожної комбінації кандидата і посилання було побудовано і оцінено 30 моделей. В результаті були відібрані моделі з найвищою точністю. Підготовка даних аналогічна раніше згаданим установкам, набір даних зменшений на 5% за рахунок видалення дублікатів, а атрибути нормалізовані для виявлення викидів.

Таблиця 4.7 – Огляд кращих результатів використання k-найближчих сусідів для відповідних комбінацій кандидата і еталона

k- найближчих сусідів					
кандидат	посилання 1	посилання 2	точність	F1-Автоматизований	F1-професійний
Google	Bing	- -	67.43%	0.763	0.631
Google	вільний переклад	- -	69.54%	0.701	0.681
Google	Bing	вільний переклад	**72.43%**	0.712	0.643
Bing	Google	- -	71.73%	0.723	0.685
Bing	вільний переклад	- -	68.35%	0.717	0.692
Bing	Google	вільний переклад	71.63%	0.728	0.673

У наступній таблиці представлені результати оцінок з використанням алгоритму наївний Байес. З огляду на характер алгоритму, результати є остаточними для набору даних та не підлягають подальшій оптимізації. Набір даних аналогічний попереднім результатам: Всі доступні атрибути нормалізуються і використовуються. Набір даних зменшений до 5% викидів і не містить дублікатів.

Таблиця 4.8 – Огляд результатів використання наївного Байеса для відповідних комбінацій кандидат-референт

Наївний Байес					
кандидат	посилання 1	посилання 2	точність	F1-автоматизований	F1-професійний
Google	Bing	- -	66.65%	0.674	0.647
Google	вільний переклад	- -	67.84%	0.673	0.673
Google	Bing	вільний переклад	**69.46%**	0.692	0.704
Bing	Google	- -	62.64%	0.683	0.684
Bing	вільний переклад	- -	67.74%	0.674	0.705
Bing	Google	вільний переклад	65.73%	0.669	0.684

Найбільша точність, досягнута для методу головних компонент, показана в таблиці 4.9. Використовуваний набір даних і атрибути аналогічні раніше показаним алгоритмам на основі речень.

Таблиця 4.9 – Огляд кращих результатів використання методу головних компонент для відповідних комбінацій кандидат-референт

SVM з підтримкою					
кандидат	посилання 1	посилання 2	точність	F1-автоматизований	F1-професійний
Google	Bing	- -	**64.52%**	0.638	0.638
Google	вільний переклад	- -	62.52%	0.693	0.638
Google	Bing	вільний переклад	63.74%	0.592	0.613
Bing	Google	- -	61.04%	0.614	0.559
Bing	вільний переклад	- -	60.63%	0.583	0.621
Bing	Google	вільний переклад	60.27%	0.672	0.603

Машини з опорних векторів були оптимізовані 30 разів. У таблиці 4.10 представлені додаткові помітні результати, які з'явилися в процесі оптимізації.

Таблиця 4.10 – Огляд найпомітніших результатів по першому питанню дослідження

Повнота і Точність		
	кандидат Google	кандидат Бінг
Середнє значення Автоматизований Повнота	80.01%	73.63%
Середній професіонал за Повнота	60.38%	67.38%
Різниця у Повнота	19.63%	6.25%
Середня точність Професіонал	73.61%	74.63%
Середня точність Автоматизований	67.04%	68.72%
Різниця в точності	6.57%	5.91%

4.2 Результати, засновані на документах

Щоб оцінити не тільки речення, але й цілі документи, підхід, заснований на реченнях, застосовується на рівні документа. Як згадувалося, в повний набір даних був витягнутий з 15 технічних документів. Результати класифікації для цих 15 оригінальних документів були створені за допомогою індивідуально оптимізованого дерева рішень і представлені в наступній таблиці.

Більш узагальнений підхід, заснований на випадково виготовлених документах, використаних для оцінок, підтверджує початкові висновки. В оцінці використовується окремо оптимізована модель дерева рішень для кожної довжини документа. Кожна використана модель являє собою оптимальний результат з набору 423 протестованих Дерев рішень. Більш детальні результати представлені в таблиці 4.12.

Таблиця 4.11 – Класифікація оригінальних документів зі знанням перекладу

	Професійно перекладені		**Автоматично перекладені**	
Довжина документа	Прогноз професійний	Прогнозований клас	Прогнозування автоматизоване	Прогнозований клас
354	62.48%	1	64.75%	0
543	65.23%	1	63.74%	0
753	66.38%	1	72.24%	0
784	61.48%	1	75.47%	0
751	69.32%	1	71.36%	0
1322	63.83%	1	72.42%	0
1234	68.73%	1	76.83%	0
1643	58.72%	1	75.19%	0
1732	69.32%	1	67.51%	0
1789	66.73%	1	73.62%	0
2476	65.62%	1	65.19%	0
3277	58.32%	1	70.51%	0
3756	71.03%	1	72.81%	0
3183	71.73%	1	68.42%	0

Таблиця 4.12 – Результати класифікації документів з урахуванням всіх атрибутів

Оцінки документів						
Довжина документа	Кількість документів	Клас	Середній% автоматизований	Середній% професіонал	Помилки в класифікації	Помилки в класифікації
5	5000	1	33.45%	68.54%	798	12.45%
10	2500	1	37.47%	69.37%	120	8.32%
25	1250	1	33.73%	69.65%	49	2.18%
50	500	1	37.37%	69.05%	5	0.71%
100	250	1	32.73%	69.29%	2	0.20%
250	50	1	34.82%	69.04%	1	0.00%
500	25	1	34.42%	70.18%	0	0.00%
1500	10	1	37.31%	69.38%	0	0.00%
3000	5	1	34.27%	69.45%	0	0.00%
5	5000	0	74.64%	32.73%	875	29.63%
10	2500	0	71.62%	31.09%	321	13.09%
25	1250	0	67.43%	31.28%	52	7.45%

Продовження таблиці 4.12

50	500	0	71.72%	28.38%	6	4.40%
100	250	0	70.09%	27.41%	1	1.30%
250	50	0	71.71%	27.19%	1	0.40%
500	25	0	72.32%	30.32%	0	0.00%
1500	10	0	71.38%	29.12%	0	0.00%
3000	5	0	68.82%	31.71%	0	0.00%

4.3 Класифікація на основі пропозицій зі знанням вихідного документа

Найкращі результати серед всіх алгоритмів для перших питань дослідження були досягнуті штучною нейронною мережею з точністю 73,25%. Однак три з п'яти протестованих алгоритмів не мають істотних відхилень один від одного, коли мова йде про точність. Середня точність трьох алгоритмів відрізняється менш ніж на два процентних пункти за всіма протестованими наборами даних, при цьому нейронна мережа досягла найвищої середньої точності - 71,52%. Що стосується відхилень від середніх значень, дерева рішень і штучні нейронні мережі відхиляються від своїх середніх значень на однакову величину (σ = 0,016), при цьому класифікатор k-найближчих сусідів досяг мінімального відхилення серед всіх алгоритмів з σ = 0,008.

Додатковим предметом інтересу є порівняння результатів, отриманих з використанням одного посилання, і результатів, отриманих з використанням декількох посилань. Основним моментом є те, що додаткове посилання не обов'язково призводить до поліпшення точності алгоритмів. Це підтверджується наведеними результатами, оскільки найкраща нейронна мережа використовує в якості еталону одну систему машинного перекладу. З іншого боку, другий кращий алгоритм використовує два алгоритми , що дозволяє припустити, що досягнута точність алгоритму частково не залежить від кількості використовуваних посилань. На відміну від багатьох підходів до оцінки систем машинного перекладу, представлений підхід намагається класифікувати задані фрагменти тексту на два класи, а не оцінювати їх якість. Оскільки багато з використовуваних атрибутів

розраховують бали подібності між еталоном і перекладом-кандидатом, еталон машинного перекладу може використовуватися для ідентифікації автоматизованих перекладів завдяки високій подібності з даними еталона, в той час як професійні переклади можуть сильніше відхилятися від нього.

Додавання перекладу "туди-назад" для використання в якості додаткового посилання призвело до ще більшого поліпшення точності. Додавання додаткових атрибутів до бази даних в цілому корисно для алгоритму, тому що отримані результати були очікуваними.

Система Freetranslation була використана в якості еталону, оскільки використання системи перекладу "туди-назад" в якості кандидата призвело б до зміщення алгоритму, і Freetranslation сягла найменшої точності на перших етапах оптимізації. Тому поєднання двох сильніших систем переказу призводить до кращих результатів.

Підводячи підсумки, можна сказати, що найкращі результати показали точність класифікації 73,25%, що дає виграш в порівнянні з випадковим класифікатором 27,73% при використанні Bing Translator як кандидата і Google Translate як псевдореференса.

Отримані результати показують значне поліпшення в порівнянні з заданим еталоном на рівні речення. Об'єднання помічених речень для класифікації на рівні документів є правильною стратегією, оскільки використовується та ж інформація, яка була б прийнята до уваги алгоритмом, що класифікують безпосередньо на корпусі документів. Жоден з представлених алгоритмів машинного навчання не навчений на даних, які використовувалися для тестування в одній і тій же ітерації процесу, що підкріплює достовірність даних.

Висновки до розділу

В результаті цієї роботи були створені дві системи класифікації. Перша, маючи доступ до оригіналу документа, здатна передбачити, чи був текст

переведений автоматично з точністю 71,84% на рівні речень і з правильними прогнозами на рівні документів при обсязі понад 50 пропозицій.

Друга система класифікації, яка не має доступу до вихідного документу, досягає 63,51% правильної класифікації на рівні речень і правильних передбачень на рівні речень. На рівні документа, враховуючи розмір 200 або більше речень в документі.

Створено структуру, яка дозволяє ранжувати перекладені фрагменти тексту на чотири класи, на основі комбінації двох алгоритмів і зваженої оцінки помилок.

Для майбутніх досліджень представлена база даних перекладів, що містить 19456 речень, включаючи контекстну інформацію і їх професійні переклади.

Загальні висновки

Розроблено метод класифікації документів за допомогою процесу виявлення знань, що складається з етапів: збір даних, попередня обробка даних, вибір відповідного підходу до пошуку закономірностей в даних і їх інтерпретація. База даних документів була розбита на рівні речень, в результаті чого було отримано дев'ять наборів даних. Були обрані та реалізовані метрики і атрибути, з яких 13 потребують еталонного перекладу для розрахунку, а 15 - ні. Для створення еталонного перекладу використовувалися одна або дві системи комп'ютерного перекладу, відповідно, для перекладу вихідного документа і створення еталона для заданих текстів-кандидатів. Описаний набір даних був попередньо оброблений, видалені 5% викидів і атрибутів, що корелюють один з одним більш ніж на 90%, а також нормалізовані дані для отримання порівнянних значень атрибутів. Попередньо оброблені дані були використані в декількох ітераціях для п'яти алгоритмів машинного навчання, Дерев рішень, Штучних нейронних мереж, k-найближчих сусідів, наївний Байес і метод головних компонент. Алгоритми були оптимізовані і протестовані на проміжній множині, яка була виділена з бази даних перед навчанням моделей. Максимальних результатів домігся класифікатор k-найближчих сусідів, який набрав 73,25% при наявності доступу до оригінального документу і 63,74% без доступу до нього. Щоб зробити висновок про класифікацію на рівні документа, оптимізовані алгоритми були використані на реченнях кожного оригінального документа, а результати були повторно об'єднані для класифікації відповідної документації. Наявність доступу до оригінального тексту не привело до помилок в класифікації 15 документів, в той час як відсутність доступу до нього показало, що коефіцієнт помилкової класифікації склав 13,81%.

Крім того, була розроблена система оцінки для ранжирування речень і документів на основі їх якості незалежно від типу перекладу. Запропонована модель складається з чотирьох класів, які використовують дві оптимізовані моделі машинного навчання для класифікації речень і додатковий незалежний від посилань

інструмент перевірки граматики і орфографії для створення виваженої кількості помилок для кожного речення. Для оцінки якості документа класи якості відповідних речень усереднюються з додатковою вагою для зменшення кількості помилок.

Перелік посилань

1. Young T. et al. Recent trends in deep learning based natural language processing //ieee Computational intelligenCe magazine. – 2018. – Т. 13. – №. 3. – С. 55-75.

2. Lavie, A., Denkowski, M.J. The Meteor metric for automatic evaluation of machine translation. *Machine Translation* **23,** 105–115 (2009). https://doi.org/10.1007/s10590-009-9059-4

3. Guzmán F. et al. Machine translation evaluation with neural networks //Computer Speech & Language. – 2017. – Т. 45. – С. 180-200.

4. Yuan Y., Sharoff S. Investigating the Influence of Bilingual MWU on Trainee Translation Quality //Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). – 2018.

5. Jiang J., Xu J., Lin Y. (2011) A Naïve Automatic MT Evaluation Method without Reference Translations. In: Wang Y., Li T. (eds) Knowledge Engineering and Management. Advances in Intelligent and Soft Computing, vol 123. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-25661-5_62

6. Comelles E., Atserias J. VERTa: a linguistic approach to automatic machine translation evaluation //Language Resources and Evaluation. – 2019. – Т. 53. – №. 1. – С. 57-86.

7. Albrecht, Joshua S., and Rebecca Hwa. “Regression for Machine Translation Evaluation at the Sentence Level.” Machine Translation, vol. 22, no. 1/2, Springer, 2008, pp. 1–27, http://www.jstor.org/stable/40285150.

8. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. AI Magazine, 17(3), 37. https://doi.org/10.1609/aimag.v17i3.1230

9. Pazmiño-Maji R. A., García-Peñalvo F. J., Conde-González M. Á. Statistical implicative analysis approximation to KDD and data mining: A systematic and mapping review in knowledge discovery database framework. – 2017.

10. Ferreira R. P. et al. Knowledge discovery in database of labor absenteeism using computational intelligence/descoberta de conhecimento em base de dados de absenteismo trabalhista com uso de inteligencia computacional/descubrimiento de conocimiento en base de datos de absentismo laboral utilizando inteligencia computacional //Gestao & Tecnologia. – 2020. – Т. 20. – №. 4. – С. 108-136.

11. Nichols J. A., Chan H. W. H., Baker M. A. B. Machine learning: applications of artificial intelligence to imaging and diagnosis //Biophysical reviews. – 2019. – Т. 11. – №. 1. – С. 111-118.

12. Camgoz N. C. et al. Neural sign language translation //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. – 2018. – С. 7784-7793.

13. Van Biljon E., Pretorius A., Kreutzer J. On optimal transformer depth for low-resource language translation //arXiv preprint arXiv:2004.04418. – 2020.

14. Damanik I. S. et al. Decision tree optimization in C4. 5 algorithm using genetic algorithm //Journal of Physics: Conference Series. – IOP Publishing, 2019. – Т. 1255. – №. 1. – С. 012012.

15. Cherfi A., Nouira K., Ferchichi A. Very fast C4. 5 decision tree algorithm //Applied Artificial Intelligence. – 2018. – Т. 32. – №. 2. – С. 119-137.

16. Phu V. N. et al. A decision tree using ID3 algorithm for English semantic analysis //International Journal of Speech Technology. – 2017. – Т. 20. – №. 3. – С. 593-613.

17. D. Y. Singh and A. S. Chauhan, “Neural networks in data mining,” Journal of Theoretical and Applied Information Technology, vol. 5, no. 1, pp. 37–42, 2009.

18. Júnior C. C. et al. Artificial Neural Networks and Data Mining Techniques for Summer Crop Discrimination: A New Approach //Canadian Journal of Remote Sensing. – 2019. – Т. 45. – №. 1. – С. 16-25.

19. Abdalla M. I. Neural Networks Based Data Mining Techniques (Dept. E) //MEJ. Mansoura Engineering Journal. – 2020. – Т. 32. – №. 4. – С. 65-75.

20. Chen S. et al. A novel selective naïve Bayes algorithm //Knowledge-Based Systems. – 2020. – Т. 192. – С. 105361.

21. Berrar D. Bayes' theorem and naive Bayes classifier //Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics; Elsevier Science Publisher: Amsterdam, The Netherlands. – 2018. – С. 403-412.

22. Jiang L. et al. Class-specific attribute weighted naive Bayes //Pattern recognition. – 2019. – Т. 88. – С. 321-330.

23. Geler Z. et al. Weighted kNN and constrained elastic distances for time-series classification //Expert Systems with Applications. – 2020. – Т. 162. – С. 113829.

24. Shah K. et al. A comparative analysis of logistic regression, random forest and KNN models for the text classification //Augmented Human Research. – 2020. – Т. 5. – №. 1. – С. 1-16.

25. Chauhan V. K., Dahiya K., Sharma A. Problem formulations and solvers in linear SVM: a review //Artificial Intelligence Review. – 2019. – Т. 52. – №. 2. – С. 803-855.

26. Gopi A. P. et al. Classification of tweets data based on polarity using improved RBF kernel of SVM //International Journal of Information Technology. – 2020. – С. 1-16.

27. Singh S. P. et al. Machine translation using deep learning: An overview //2017 international conference on computer, communications and electronics (comptelix). – IEEE, 2017. – С. 162-167.

28. Poibeau T. Machine translation. – MIT Press, 2017.

Ministry of Education and Science of Ukraine

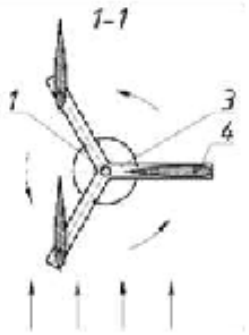
Khmelnytskyi National University

Ukrainian-Polish Scientific Dialogues

International Conference

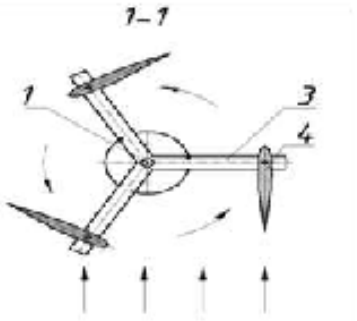
20 - 23 October 2021

Khmelnytskyi – Kamianets-Podilskyi



Напрямок потоку повітря

Рис.2 Горизонтальний переріз 1-1



Напрямок потоку повітря

Рис. 4 Горизонтальний переріз Б-Б

Рис. 4 Схема сил що діють на лопать

Оскільки запропонована конструкція ВЕУ дозволяє генерувати енергію при малих швидкостях вітру, то це дає можливість отримати більшу кількість енергії вітру, враховуючи, що на території України середня швидкість вітру коливається в районі 2…4 м/с.

Література:

1. Д. Н. Горелов Полуэмпирический метод расчета оптимальных геометрических параметров ротора Дарье / Прикладная механика и техническая физика. 2015. т. 56, № 3. 91-104 с.

2. Вітроенергетична установка в вертикальним ротором: пат. 136289 Україна : F03D 3/00. № u201902252; заявл. 12.08.2019; опубл. 12.08.2019, Бюл. № 15. 5 с.

СКРИПНИК Т., МАНЗЮК Е. .[1]

[1] Хмельницький національний університет, Україна

АНАЛІТИЧНА СИСТЕМА ВИЗНАЧЕННЯ ЯКОСТІ ПЕРЕКЛАДУ ТЕКСТОВОЇ ІНФОРМАЦІЇ МЕТОДАМИ МАШИННОГО НАВЧАННЯ

Розроблено на реалізовано систему оцінки якості перекладу технічної документації за допомогою методів машинного навчання, які базуються на сукупності ознак та виявленню відмінностей між автоматизованим перекладом та професійним перекладом, який виконано людиною. Було сформовано множину ознак, як вхідних параметрів якості перекладу для методі машинного навчання. Використовувались методи з наявністю еталону перекладу для порівняння та без нього. Результати дослідження показали прийнятні оцінки ефективності запропонованих методів.

A system for assessing the quality of translation of technical documentation using machine learning methods, which are based on a set of features and identifying differences between machine

translation and professional translation performed by humans, has been developed and implemented. Many features were formed as input parameters of translation quality for machine learning methods. Methods with and without a translation standard were used for comparison. The results of the study showed acceptable evaluations of the effectiveness of the proposed methods.

В умовах все поширення інформації наявність високоякісних перекладів має вирішальне значення для успіху в умовах зростаючої міжнародної конкуренції. Великі міжнародні компанії, а також компанії середнього розміру повинні надавати своїм клієнтам добре перекладену технічну документацію високої якості не тільки для того, щоб бути успішними на ринку, але і для того, щоб відповідати правовим нормам і уникнути судових позовів.

Технічна документація - це загальний термін для кожного виду документа, пов'язаного з продукцією, метою якого є розкриття інформації про продукт або послугу. Технічна документація поділяється на внутрішню і зовнішню [1]. Під внутрішньою документацією розуміються технічні креслення, переліки деталей, переліки робіт, робочі інструкції та ін. Вона є основоположною при розробці, створенні та обслуговуванні продукції. Зовнішня документація, що включає технічні паспорти, каталоги запасних частин і керівництва, адресована існуючим клієнтам і частково використовується для їх придбання [2, 3]. Зовнішня документація також підтверджує специфікації продукції для державних органів. Різні види і кілька функцій вимагають наявності технічної документації для виконання певних вимог, наприклад:

- Аудиторія повинна бути відома і відповідним чином адресована.
- Мовний стиль повинен бути спрямований на розуміння описуваного питання.
- Документація повинна бути повною і структурованої відповідно до потреб користувачів.
- Необхідно дотримуватися законів і стандарти.
- Документ повинен бути привабливим і як можна більш лаконічним.

Всі технічні документи підкреслюють зрозумілість як ключовий аспект своєї структури. Головна мета технічної документації полягає в тому, щоб кінцеві користувачі, а також співробітники різних відділів компанії могли зрозуміти її без додаткових досліджень. Це гарантує, що тексти в основному написані ясно, просто і лаконічно і не містять занадто багато скорочень або внутрішньої корпоративної термінології. Крім того, технічна документація зазвичай багаторазово перевіряється перед публікацією або для забезпечення її якості за допомогою коректорів, або для забезпечення її зрозумілості за допомогою аналізу аудиторії.

Метою даної роботи була оцінка якості технічних документів та їх перекладів за допомогою методів машинного навчання з акцентом на виявлення відмінностей між автоматичними перекладами і професійними перекладами, виконаними людьми.

Для аналізу фрагментів тексту кандидата на предмет їх словесної і граматичної правильності була використана система перевірки стилістичних слів і граматики без виявлення семантичної складової тексту Програма підраховує всі знайдені помилки і розділяє їх на різ-

ні категорії. На додаток до загальної кількості помилок були додані окремі категорії, в результаті чого було отримано ще 72 ознаки.

Можливості оцінки якості документа без знання вихідного тексту обмежені. На відміну від цього, продуктивність і якість передбачення алгоритму машинного навчання сильно залежать від доступних атрибутів для навчання і тестування. Це є проблемою для вирішення даної частини питання дослідження, яка може поставити під загрозу його успіх. Для вирішення цієї проблеми було створено еталон "прямий-зворотній". Це дозволило створити додаткові атрибути, згадані вище, без використання вихідного документа, в результаті чого з'явилися ще атрибути, які могли бути використані алгоритмами машинного навчання.

В результаті цієї роботи були створені дві системи класифікації. Перша, маючи доступ до оригіналу документа, здатна передбачити, чи був текст переведений автоматичному з точністю 67% на рівні пропозицій і з правильними прогнозами на рівні документів при обсязі понад ста пропозицій.

Друга система класифікації, яка не має доступу до вихідного документу, досягає 58% правильної класифікації на рівні пропозицій і правильних передбачень на рівні пропозицій.

Перелік посилань:

1. Caponi A. et al. Exploiting patterns and templates for technical documentation //Proceedings of the ACM Symposium on Document Engineering 2018. – 2018. – С. 1-9.
2. Rydén A. et al. Technical Documentation. – 2020.
3. Baratov D., Astanaliev E. Functional Features of the Technical Documentation Control Program //International Journal on Orange Technologies. – 2021. – Т. 3. – №. 1. – С. 7-11.

СТАРИЙ А., ГОРДЄЄВ А. .[1]

[1] Хмельницький національний університет, Україна

ЕКСПЕРИМЕНТАЛЬНЕ ДОСЛІДЖЕННЯ ПРОЦЕСУ МИЙКИ ЗАБРУДЕННЯ ПУЛЬСУЮЧИМ СТРУМЕНЕМ РІДИНИ З КАВІТАЦІЙНИМИ ПУХИРЦЯМИ

Experimental studies of the process of washing pollution by pulsating flow of liquid with cavitation bubbles

Studies of model contamination washing confirmed the main theoretical assumptions on the mechanical nature of the interaction of detergent with contamination and showed the effectiveness of the method of washing with cavitation bubbles

На поверхні деталей і складальних вузлів в процесі їх виготовлення, експлуатації машин і устаткування утворюються технологічні та виробничі забруднення. При технічному обслуговуванні та ремонті виникає необхідність мийки деталей при їх збиранні у вузли. Режими мийки поверхні деталі потоком миючої рідини визначають, виходячи з аналізу гідродинамічної взаємодії миючої рідини з частинками з існуючим забрудненням на деталях, а також на підставі аналізу результатів експериментальних досліджень. В процесі мийки в основному проходить зрив частинок забруднення потоком миючої рідини завдяки дії нормаль-

CERTIFICATE

20 - 23 October 2021

Khmelnytskyi – Kamianets-Podilskyi

ISSUED TO CERTIFY THAT

Скрипник Т.

participated in the international conference
"IX Ukrainian-Polish Scientific Dialogues"(24 hours/0,8 ECTS credits)

Serhii Matiukh
rector

signature

Khmelnytskyi National University

КВАЛІФІКАЦІЙНА РОБОТА МАГІСТРА

МЕТОД КЛАСИФІКАЦІЇ ТЕКСТОВИХ ДОКУМЕНТІВ ЗАСОБАМИ МАШИННОГО НАВЧАННЯ

ВИКОНАЛА:
СТУДЕНТКА 2 КУРСУ, ГРУПИ КНм-20-2
СКРИПНИК ТЕТЯНА КАЗИМИРІВНА

КЕРІВНИК:
Д.Т.Н., ПРОФЕСОР КАФЕДРИ КН
БАРМАК ОЛЕКСАНДР ВОЛОДИМИРОВИЧ

Актуальність теми

В магістерській роботі розроблено та набуло практичної реалізації метод класифікації технічної документації на основі пропозиції перекладу.

Оцінка перекладеної технічної документації є важливим кроком для компаній, що дозволяє скоротити час і витрати, а також створити ефективний спосіб перекладу важливих документів. Крім того, це забезпечує певний рівень якості.

Метою дослідження є розробка методу реалізації процесу машинного навчання, який здатний класифікувати текстові дані, та визначити чи були документи переведені професійними перекладачами або комп'ютерними системами трансляції текстової інформації.

Для досягнення зазначеної мети поставлені наступні **задачі**:

- визначити мету процесу і зібрати попередні необхідні знання про прикладну область;
- вибрати відповідний набір даних для отримання знань;
- попередня обробка даних;
-привести дані в прийнятний формат;
-прийняти рішення про підхід до отримання даних для певної мети процесу отримання знань;
-вибір алгоритму аналізу даних;

- основний етап отримання даних, який полягає в застосуванні алгоритму до попередньо обробленого набору даних та пошуку цінних знань у даних;
- інтерпретувати знайдені алгоритмом закономірності і, можливо, повернутися до одного з попередніх етапів, щоб скорегувати настройку процесу отримання знань;
- використання інтерпретованих результатів для подальших дій, наприклад, для подальшого дослідження або застосування систем до реального сценарію.

Об'єктом дослідження є процеси отримування інформації з використанням методів машинного навчання.

Предметом дослідження є моделі та методи обробки текстових даних технічної документації та перекладених документів різними мовами.

Наукова новизна одержаних результатів.

В результаті проведеної роботи були отримані такі результати:

- набула подальшого розвитку система оцінки якості класифікації текстової інформації, яка специфікується областю дослідження та є інтегральним показником якості класифікації;
- запропоновано інноваційний метод визначення якості перекладу текстової технічної документації на основі класифікації перекладу виконаного перекладачами та автоматизованими системами перекладу;
- запропоновано метод визначення якості перекладу із застосуванням методів розмічених та нерозмічених даних.

Практичне значення одержаних результатів

Отримані практичні результати досліджень можуть бути застосовні для визначення якості перекладу технічної документації.

Запропонована модель складається з методів, які використовують дві оптимізовані моделі машинного навчання для класифікації пропозицій і додатковий незалежний від посилань інструмент перевірки граматики і орфографії для створення виваженої кількості помилок для кожної пропозиції.

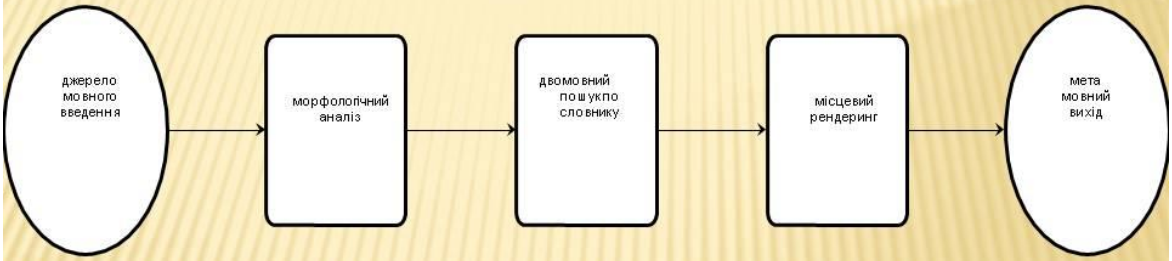
Представлена система класифікації технічних документів з використанням методів машинного навчання і підхід до оцінки якості документів.

Апробація кваліфікаційної роботи

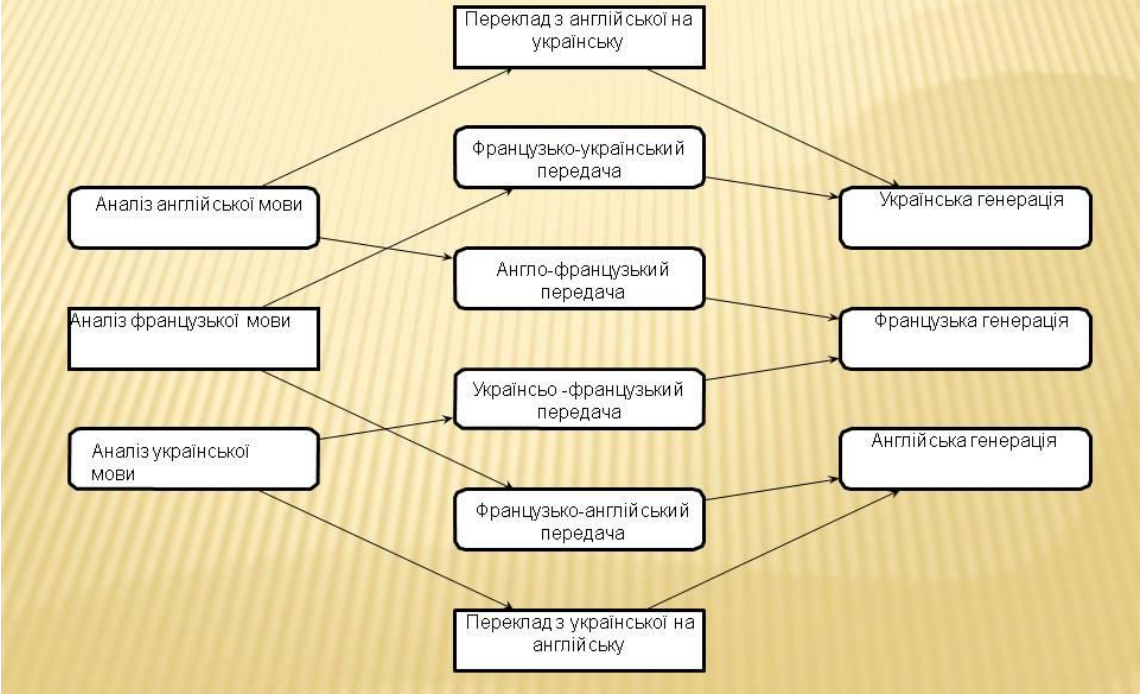
– Доповідь на тему “Аналітична система визначення якості перекладу текстової інформації методами машинного навчання” на Міжнародній конференції «IX Українсько-Польські наукові діалоги», Хмельницький, Україна 20-23 жовтня 2021 р.

– Доповідь на тему “Метод машинного навчання для визначення якості перекладу текстової інформації” на Всеукраїнській конференції АПКН-2021, Хмельницький, Україна 15 листопада 2021 р.

Процес прямого перекладу



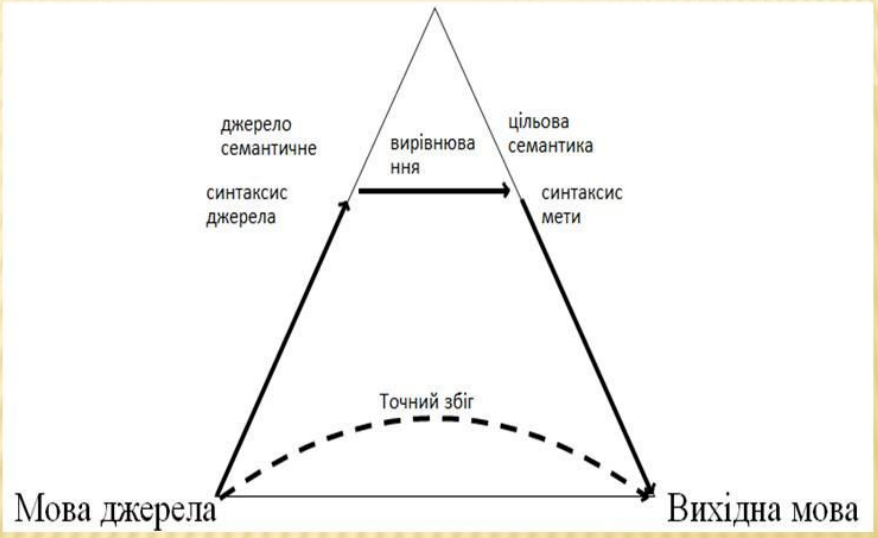
Візуалізація залежності мови та напрямки перекладу на етапі перекладу в процесі трансферного перекладу



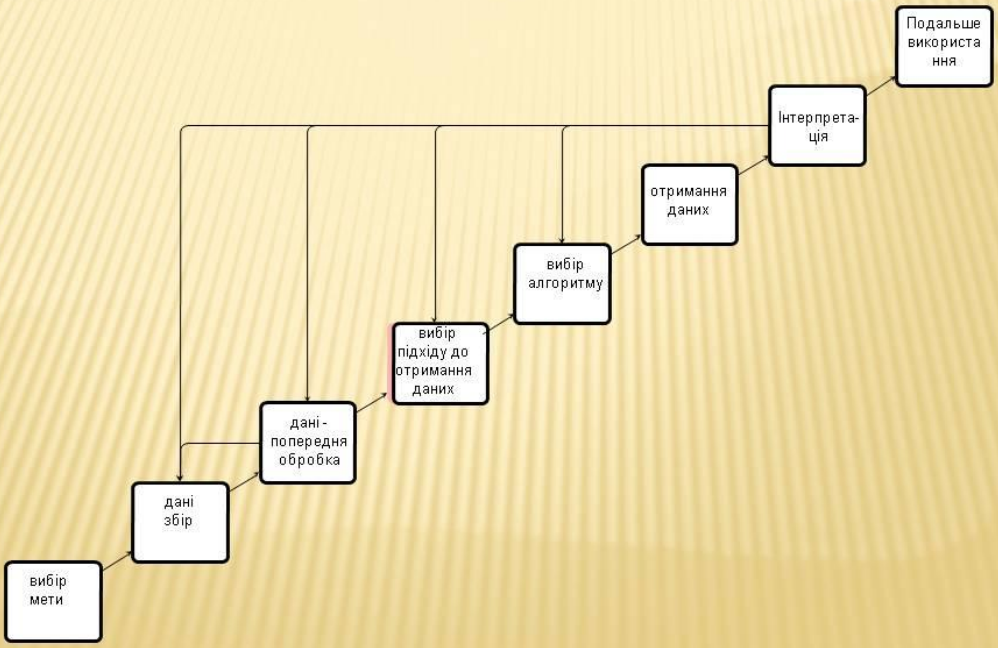
Процес перекладу на інтерлінгва для двох підтримуваних мов

Візуалізація обсягу аналізу, виконуваного в процесі машинного перекладу на основі правил

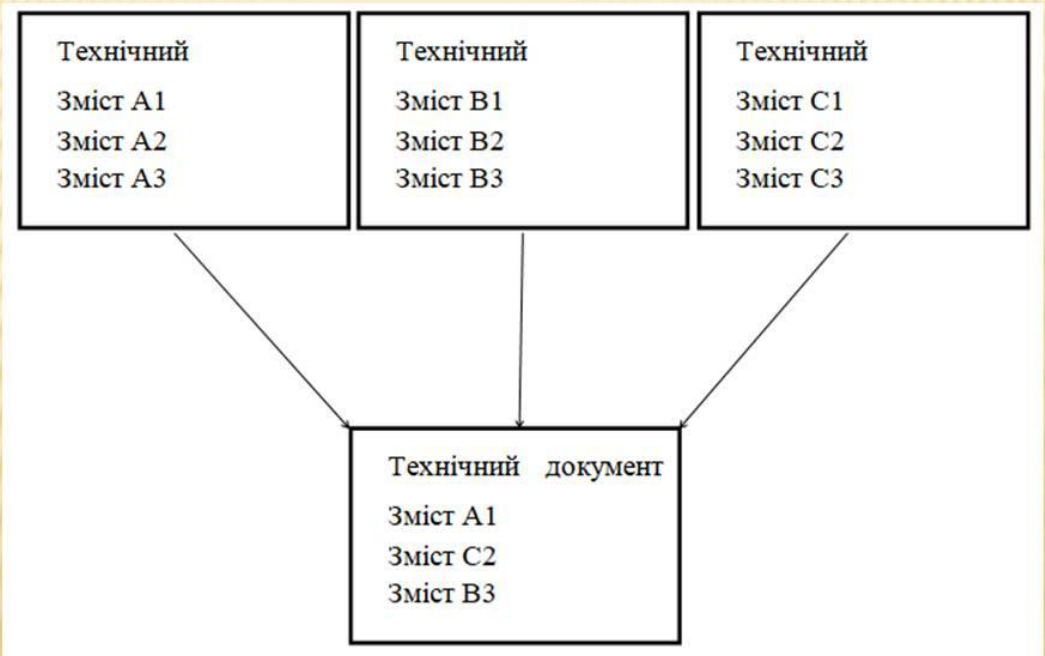
Адаптована версія трикутника



Адаптований процес виявлення знань, що містить кроки від збору даних до інтерпретації результатів і їх використання для вирішення подальших завдань



Створення технічного документа



Статистика використаних наборів даних

Кількість систем автоматизованого перекладу	3
Кількість технічних документів	15
Початкова кількість пропозицій у всіх документах	30
Кількість пропозицій, використаних кожною системою перекладу для наборів даних	327
Приблизна кількість пропозицій в наборах даних	34, 730
Загальна кількість наборів даних	7
Кількість атрибутів, необхідних для еталонного перекладу	14
Кількість атрибутів, які не потребують перекладу посилань	19
Загальна кількість атрибутів	42
Максимальна кількість доступних значень атрибутів	1 456 210
Кількість створених фіктивних документів	190

Використовувані комбінації референтів і кандидатів

Кандидат	Посилання1	Посилання2	Переклад в обидва кінці
Google Translate	Bing	- -	Google ПТН через Freetranslation
Google Translate	вільний переклад	- -	Google ПТН через Freetranslation
Google Translate	Bing	вільний переклад	Google ПТН через Freetranslation
перекладач Bing	Google	- -	Bing ПТН через Freetranslation
перекладач Bing	вільний переклад	- -	Bing ПТН через Freetranslation
перекладач Bing	Google	вільний переклад	Bing ПТН через Freetranslation
вільний переклад	Google	- -	- -
вільний переклад	Bing	- -	- -
вільний переклад	Google	Bing	- -

Огляд створених моделей і оптимізацій

	Дослідницький питання 1	Дослідницький питання 2	**всього**
Побудовані дерева рішень	320 000	200 000	680 000
Побудовані нейронні мережі	5 000	1 000	7 000
Побудова k-найближчих сусідів	3 100	1 500	54 000
Побудовані опорні SVM	2 200	1 000	2 500
Загальна кількість оптимізацій	323 000	204 500	694 500

Середні значення та стандартні відхилення протестованих алгоритмів

алгоритм	точність	Стандартне відхилення
дерево рішень	67.34%	0.013
Штучна нейронна мережа	**72.45%**	0.015
k-Nearest Neighbor	68.45%	**0.010**
наївний Байес	67.67%	0.020
SVM з підтримкою	63.76%	0.018

Огляд кращих результатів використання дерева рішень для відповідних комбінацій кандидата і еталона

Дерево рішень					
кандидат	посилання 1	посилання 2	точність	F1-автоматизований	F1-професійний
Google	Bing	- -	68.19%	0.705	0.743
Google	вільний переклад	- -	64.93%	0.758	0.683
Google	Bing	вільний переклад	**73.43%**	0.738	0.684
Bing	Google	- -	72.19%	0.782	0.662
Bing	вільний переклад	- -	67.76%	0.737	0.658
Bing	Google	вільний переклад	68.54%	0.783	0.683
вільний переклад	Bing	- -	64.21%	0.762	0.651
вільний переклад	Google	- -	63.63%	0.652	0.628
вільний переклад	Bing	Google	66.73%	0.689	0.649

Огляд кращих результатів використання штучної нейронної мережі для відповідних комбінацій кандидат-референт

Штучна нейронна мережа					
кандидат	посилання 1	посилання 2	точність	F1-автоматизо ваний	F1-професійний
Google	Bing	- -	70.43%	0.729	0.681
Google	вільний переклад	- -	66.65%	0.727	0.681
Google	Bing	вільний переклад	71.34%	0.756	0.671
Bing	Google	- -	**73.25%**	0.721	0.724
Bing	вільний переклад	- -	69.82%	0.681	0.701
Bing	Google	вільний переклад	71.84%	0.714	0.723

Огляд кращих результатів використання k-найближчих сусідів для відповідних комбінацій кандидата і еталона

k- найближчих сусідів					
кандидат	посилання 1	посилання 2	точність	F1-Автоматизов аний	F1-професійний
Google	Bing	- -	67.43%	0.763	0.631
Google	вільний переклад	- -	69.54%	0.701	0.681
Google	Bing	вільний переклад	**72.43%**	0.712	0.643
Bing	Google	- -	71.73%	0.723	0.685
Bing	вільний переклад	- -	68.35%	0.717	0.692
Bing	Google	вільний переклад	71.63%	0.728	0.673

Огляд результатів використання наївного Байеса для відповідних комбінацій кандидат-референт

Наївний Байес					
кандидат	посилання 1	посилання 2	точність	F1-автоматизований	F1-професійний
Google	Bing	- -	66.65%	0.674	0.647
Google	вільний переклад	- -	67.84%	0.673	0.673
Google	Bing	вільний переклад	**69.46%**	0.692	0.704
Bing	Google	- -	62.64%	0.683	0.684
Bing	вільний переклад	- -	67.74%	0.674	0.705
Bing	Google	вільний переклад	65.73%	0.669	0.684

Огляд кращих результатів використання методу головних компонент для відповідних комбінацій кандидат-референт

SVM з підтримкою					
кандидат	посилання 1	посилання 2	точність	F1-автоматизований	F1-професійний
Google	Bing	- -	**64.52%**	0.638	0.638
Google	вільний переклад	- -	62.52%	0.693	0.638
Google	Bing	вільний переклад	63.74%	0.592	0.613
Bing	Google	- -	61.04%	0.614	0.559
Bing	вільний переклад	- -	60.63%	0.583	0.621
Bing	Google	вільний переклад	60.27%	0.672	0.603

Огляд найпомітніших результатів по першому питанню дослідження

Повнота і Точність		
	кандидат Google	кандидат Бінг
Середнє значення Автоматизований Повнота	80.01%	73.63%
Середній професіонал за Повнота	60.38%	67.38%
Різниця у Повнота	19.63%	6.25%
Середня точність Професіонал	73.61%	74.63%
Середня точність Автоматизований	67.04%	68.72%
Різниця в точності	6.57%	5.91%

Класифікація оригінальних документів зі знанням перекладу

	Професійно перекладені документи		**Автоматично перекладені документи**	
Довжина документа	Прогноз професійний	Прогнозований клас	Прогнозування автоматизоване	Прогнозований клас
354	62.48%	1	64.75%	0
543	65.23%	1	63.74%	0
753	66.38%	1	72.24%	0
784	61.48%	1	75.47%	0
751	69.32%	1	71.36%	0
1322	63.83%	1	72.42%	0
1234	68.73%	1	76.83%	0
1643	58.72%	1	75.19%	0
1732	69.32%	1	67.51%	0
1789	66.73%	1	73.62%	0
2476	65.62%	1	65.19%	0
3277	58.32%	1	70.51%	0
3756	71.03%	1	72.81%	0
3183	71.73%	1	68.42%	0

Результати класифікації документів з урахуванням всіх атрибутів

Оцінки документів						
Довжина документа	Кількість документів	Клас	Середній% автоматизований	Середній% професіонал	Помилки в класифікації	Помилки в класифікації
5	5000	1	33.45%	68.54%	798	12.45%
10	2500	1	37.47%	69.37%	120	8.32%
25	1250	1	33.73%	69.65%	49	2.18%
50	500	1	37.37%	69.05%	5	0.71%
100	250	1	32.73%	69.29%	2	0.20%
250	50	1	34.82%	69.04%	1	0.00%
500	25	1	34.42%	70.18%	0	0.00%
1500	10	1	37.31%	69.38%	0	0.00%
3000	5	1	34.27%	69.45%	0	0.00%
5	5000	0	74.64%	32.73%	875	29.63%
10	2500	0	71.62%	31.09%	321	13.09%
25	1250	0	67.43%	31.28%	52	7.45%
50	500	0	71.72%	28.38%	6	4.40%
100	250	0	70.09%	27.41%	1	1.30%
250	50	0	71.71%	27.19%	1	0.40%
500	25	0	72.32%	30.32%	0	0.00%
1500	10	0	71.38%	29.12%	0	0.00%
3000	5	0	68.82%	31.71%	0	0.00%

ВИСНОВКИ

В результаті виконанні кваліфікаційної роботи магістра розроблено метод класифікації документів за допомогою процесу виявлення знань, що складається з етапів: збір даних, попередня обробка даних, вибір відповідного підходу до пошуку закономірностей в даних і їх інтерпретація. База даних документів була розбита на рівні пропозицій, в результаті чого було отримано дев'ять наборів даних. Були обрані та реалізовані метрики і атрибути, з яких 13 потребують еталонного перекладу для розрахунку, а 15 - ні. Для створення еталонного перекладу використовувалися одна або дві системи комп'ютерного перекладу, відповідно, для перекладу вихідного документа і створення еталона для заданих текстів-кандидатів.

Описаний набір даних був попередньо оброблений, видалені 5% викидів і атрибутів, що корелюють один з одним більш ніж на 90%, а також нормалізовані дані для отримання порівнянних значень атрибутів. Попередньо оброблені дані були використані в декількох ітераціях для п'яти алгоритмів машинного навчання, Дерев рішень, Штучних нейронних мереж, k-найближчих сусідів, наївний Байес і метод головних компонент. Алгоритми були оптимізовані і протестовані на проміжній множині, яка була виділена з бази даних перед навчанням моделей. Максимальних результатів домігся класифікатор k-найближчих сусідів, який набрав 73,25% при наявності доступу до оригінального документу і 63,74% без доступу до нього.

Щоб зробити висновок про класифікацію на рівні документа, оптимізовані алгоритми були використані на пропозиціях кожного оригінального документа, а результати були повторно об'єднані для класифікації відповідної документації. Наявність доступу до оригінального тексту не привело до помилок в класифікації 15 документів, в той час як відсутність доступу до нього показало, що коефіцієнт помилкової класифікації склав 13,81%.

Крім того, була розроблена система оцінки для ранжирування пропозицій і документів на основі їх якості незалежно від типу перекладу.

Запропонована модель складається з чотирьох класів, які використовують дві оптимізовані моделі машинного навчання для класифікації пропозицій і додатковий незалежний від посилань інструмент перевірки граматики і орфографії для створення виваженої кількості помилок для кожної пропозиції. Для оцінки якості документа класи якості відповідних пропозицій усереднюються з додатковою вагою для зменшення кількості помилок.

Tue Nov 23 10:47:23 EET 2021, Петровський Сергій Степанович, Хмельницький національний університет, ХНУ

Anti-Plagiarism v-15.257

Максимальное совпадение с одним документом 28.0%

Словари проверки: en_US, ru_RU, ua_UA. **Ошибок в документах: 3%**

ID: 97046 Название: Метод класифікації текстових документів засобами машинного навчання Добавлено в БД: 2021-11-23 Авторы: Т.К. Скрипник Руководители: О.В. Бармак Консультанты: Опоненты:	Документ		Суммарное совпадение по Базе Данных	
	Символы	Лексемы	Символы	Лексемы
	129462	920	37411 (29%)	277 (30%)

Источник плагиата

ID	Описание	Наличие плагиата в документе	
		Символы	Лексемы
95910	Название: ЗВІТ з професійної практики Добавлено в БД: 2021-09-30 Авторы: Скрипник Т.К. Руководители: Скрипник Т.К. Консультанты: Опоненты:	36361 (28.0%)	289 (31.0%)

UNICHECK

Ім'я користувача:
Кафедра КН

ID перевірки:
1009309588

Дата перевірки:
23.11.2021 11:49:23 EET

Тип перевірки:
Doc vs Internet + Library

Дата звіту:
23.11.2021 11:55:27 EET

ID користувача:
100005671

Назва документа: **Дип_Скрипник Т.К._4 Lite**

Кількість сторінок: **90** Кількість слів: **19283** Кількість символів: **148043** Розмір файлу: **1.04 MB** ID файлу: **1009334742**

2.64%
Схожість

Найбільша схожість: **1.66%** з джерелом з Бібліотеки (ID файлу: **1009334638**)

1.07% Джерела з Інтернету	126	Сторінка 92
1.7% Джерела з Бібліотеки	58	Сторінка 93

0% Цитат

Вилучення цитат вимкнене

Вилучення списку бібліографічних посилань вимкнене

0%
Вилучень

Немає вилучених джерел

Модифікації

Виявлено модифікації тексту. Детальна інформація доступна в онлайн-звіті.

Замінені символи	11

РІШЕННЯ ЕКСПЕРНОЇ КОМІСІЇ

КАФЕДРИ КОМП'ЮТЕРНИХ НАУК

ПРО ДОПУСК КВАЛІФІКАЦІЙНОЇ РОБОТИ МАГІСТРА ДО ЗАХИСТУ ЗА РЕЗУЛЬТАТАМИ АНАЛІЗУ ЗВІТУ ПОДІБНОСТІ

Підтверджуємо ознайомлення з результатом звіту подібності щодо роботи, генерованого системою виявлення текстових збігів/ідентичності/схожості:

Назва: *Метод класифікації текстових документів засобами машинного навчання*

Автор: *Скрипник Тетяна Казимирівна*

Спеціальність: 122 – Компютерні науки

Освітня програма: освітньо-професійна

Науковий керівник: д.т.н., проф. Олександр Бармак

Після аналізу звіту подібності зроблено такий висновок:

№	Висновок	Позначка про відповідність
1	Запозичення, виявлені в роботі, є законними і не є плагіатом. Робота приймається до захисту.	відповідає
2	Виявлені запозичення не є плагіатом, розміщені в розділах, які не описують безпосередньо авторське дослідження, але кількість цитат перевищує обсяг, виправданий поставленою метою роботи Робота приймається до захисту, але має бути відкоригована. Відкоригований варіант має бути поданий на кафедру за 2 дні до захисту, разом із заявою щодо самостійності виконання письмової роботи та ідентичності друкованої та електронної версії роботи	
3	Виявлені запозичення не є плагіатом, але частково розміщені в розділах, які описують безпосередньо авторське дослідження, а кількість цитат перевищує обсяг, виправданий поставленою метою роботи. В зв'язку з цим мета роботи та поставлені завдання не були досягнені. Робота може бути допущена до захисту (наступного року) після того як буде відкоригована та допрацьована і успішно пройде повторну перевірку на академічний плагіат.	
4	Робота містить навмисні текстові спотворення, передбачувані спроби укриття запозичень або інші прояви академічного плагіату. Робота містить фабрикацію або фальсифікацію даних. Робота не допускається до захисту.	

Підтвердження:

Запозичення, виявлені в роботі, є законними і не є плагіатом, оскільки:

1) за програмою Anti-Plagiarism виявлені 28% запозичень вказують на документ автора роботи Скрипник Т.К. та містять ЗВІТ з науково-дослідної практики.
2) За програмою UNICHECK виявлені 2.6% є фрагментарними – містять поширені конструкції, загальновідомі терміни, скорочення та визначення

Сумарний обсяг всіх запозичень, визначений системою виявлення збігів/ідентичності/схожості, складає 28% і 2.6% відповідно, що, з урахуванням наведених обґрунтувань, відповідає характеру наукового дослідження і свідчить на користь кваліфікаційної роботи.

Керівник роботи _____ Олександр Бармак

Гарант ОП _____ Руслан Багрій

Завідувач кафедри КН _____ Олександр Бармак

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
МОН УКРАЇНИ

кафедра комп'ютерних наук

ВІДГУК ОПОНЕНТА

на кваліфікаційну роботу магістра

гр. КНм-20-2 Скрипник Тетяна Казимирівна за темою: *Метод класифікації текстових документів засобами машинного навчання*

1. Актуальність обраної теми

Тема кваліфікаційної робота є актуальною та належним чином обґрунтована. Стосується питання визначення якості перекладу на основі текстових даних, що є актуальним завдання.

2. Відповідність роботи предметній області спеціальності 122 Комп'ютерні науки та загальним вимогам до наукових робіт

Обрана тема в межах якої реалізовані поставлені задачі повною мірою відповідає предметній області спеціальності 122 Комп'ютерні науки та вимогам до кваліфікаційної роботи магістра.

3. Повнота розкриття мети та завдань дослідження

Поставлені завдання дослідження повністю розкривають мету дослідження та поставленні в межах теми.

4. Наявність наукової новизни

В кваліфікаційній роботі представлена наукова новизна в межах обраної області дослідження. Продемонстровано та обґрунтовано результати, які мають наукове значення. Результати дослідження оприлюдненні на науковій конференції.

5. Зміст кожного розділу роботи

Робота містить чотири розділи. В першому обґрунтовано актуальність та поставлені задачі дослідження. Другий розділ присвячено розробці моделей оцінки методу класифікації. У третьому розділі представлена розробка системи класифікація текстових документів за формальними претендентами. У четвертому розділі представлено дослідження ефективності методів класифікації текстових пропозицій за формальними методами.

6. Ступінь розкриття теми роботи

Тема роботи в повній мірі обґрунтована, проведено аналіз актуальності та відомих досліджень в межах обраної теми, поставлені завдання, які реалізовані та досліджена ефективність запропонованих методів.

7. Якість оформлення кваліфікаційної роботи

Оформлення роботи відповідає необхідним нормам та вимогам, які ставлять до оформлення кваліфікаційної роботи

8. Недоліки кваліфікаційної роботи

Робота має певні недоліки які полягають у необхідності розширення корпусу експериментальних даних.

9. Загальний висновок (допускається чи не допускається до захисту), якої оцінки заслуговує кваліфікаційна робота.

Враховуючи рівень виконання та забезпечення усіх необхідних вимог робота може бути допущена до захисту. Рекомендована оцінка «відміно».

Опонент _____ д.т.н., доц. Сергій Лисенко

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
МОН УКРАЇНИ
кафедра комп'ютерних наук

ВІДГУК НАУКОВОГО КЕРІВНИКА

на кваліфікаційну роботу магістра

гр. КНм-20-2 Скрипник Тетяна Казимирівна за темою: *Метод класифікації текстових документів засобами машинного навчання*

1. Актуальність теми

В магістерській роботі був розроблений та набув практичної реалізації метод класифікації технічної документації за пропозиціями перекладу. Запропоновано метод визначення якості перекладу технічної документації за класифікацією текстових даних, які визначаються класами якісно обробленої інформації. В даний час компанії вирішують проблему перекладу технічної документації, передаючи цю задачу зовнішнім перекладачам. Оскільки особа, що замовляє такий переклад, не обов'язково володіє мовою перекладу, важливо переконатися, що робота була виконана правильно і професійно. Робота на основі навчених моделей дозволяє визначити якість перекладу.

2. Відповідність роботи предметній області спеціальності 122 Комп'ютерні науки та загальним вимогам до наукових робіт

За змістовною та структурною складовою робота відповідає вимогам, які ставлять до кваліфікаційної роботи освітнього рівня магістра. Робота містить наукову складову та за оформленням відповідає вимогам до наукових робіт. За предметом, об'єктом, метою та методами дослідження відповідає предметній області спеціальності 122 Комп'ютерні науки

3. Професійні та особистісні якості магістранта

Під час виконання кваліфікаційної роботи магістром були продемонстровано належні знання та вміння набуті під час навчання. Продемонстровано застосування кваліфікаційних компетенцій з вирішення відповідних задач наукового напрямку в предметній області роботи. За сукупністю продемонстрованих набутих компетенцій при реалізації кваліфікаційної роботи доведена відповідність освітньому рівню магістра.

4. Ступінь самостійності під час виконання кваліфікаційної роботи

При виконання роботи магістром особисто було визначено методи реалізації поставлених задач, проведено огляд наукових досліджень з напрямку роботи. Розроблені моделі та запропоновано метод реалізації задач поставлених в межах дослідження.

Проведені експериментальні дослідження та підтверджено ефективність запропонованих методів.

5. Наукова новизна та оригінальність запропонованих підходів

В роботі представлена наукова новизна. Набула подальшого розвитку система оцінки якості класифікації текстової інформації. Запропоновано інноваційний метод визначення якості перекладу текстової технічної документації на основі класифікації перекладу виконаного перекладачами та автоматизованими системами перекладу. Запропоновано метод визначення якості перекладу із застосуванням методів розмічених та нерозмічених даних. Результати роботи оприлюдненні на науковій конференції.

6. Ступінь оволодіння методами дослідження

Магістр під час виконання продемонстрував належне володіння методами машинного навчання, методами класифікації, методами наукового дослідження з відповідним експериментальним підтвердженням. На належному рівні продемонстровано володіння методологією наукових досліджень.

7. Повнота та якість розкриття теми роботи

Тема роботи повною мірою розкрита в задачах дослідження, які успішно реалізовані.

8. Логічність, послідовність, аргументованість, літературна грамотність викладу матеріалу

Записка до кваліфікаційної роботи за структурою, логічністю викладення матеріалу, аргументованістю структурою послідовного викладення необхідною мірою відповідає стандартам.

9. Можливість практичного застосування кваліфікаційної роботи, окремих її частин

Предметна область кваліфікаційної роботи базується на ряді практичних задач, які успішно реалізовані і можуть бути безпосередньо застосовані в області визначення якості перекладу.

10. Висновок про можливість допуску кваліфікаційної роботи до захисту, на яку оцінку заслуговує робота

За сукупністю вимог робота повною мірою відповідає кваліфікаційному рівню магістра, рекомендується до захисту та заслуговує на оцінку «відміно».

Науковий керівник _____ д.т.н., проф. Олександр Бармак