ДОДАТОК А

Слайди презентації

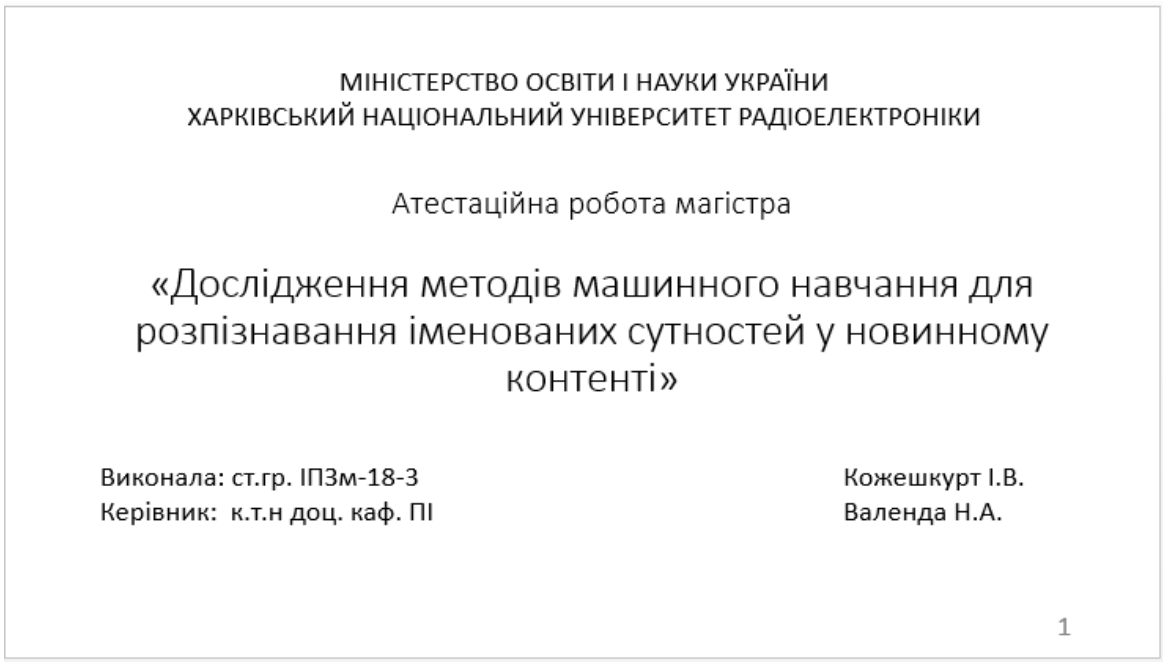


Рисунок А.1 – Слайд 1

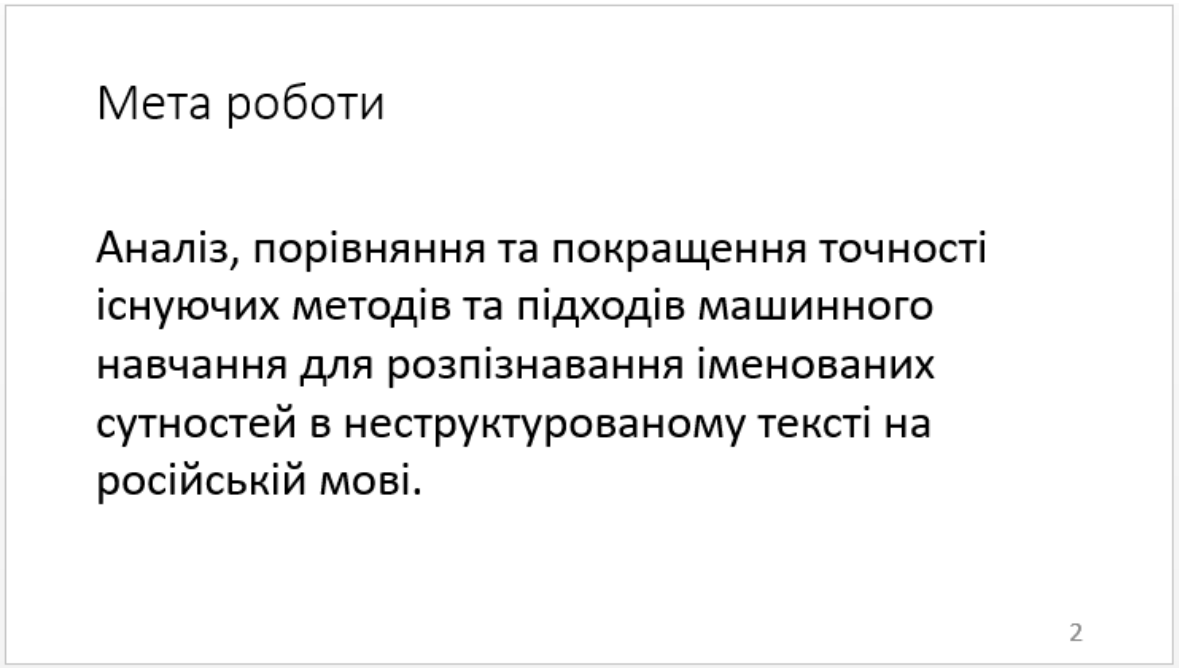


Рисунок А.2 – Слайд 2

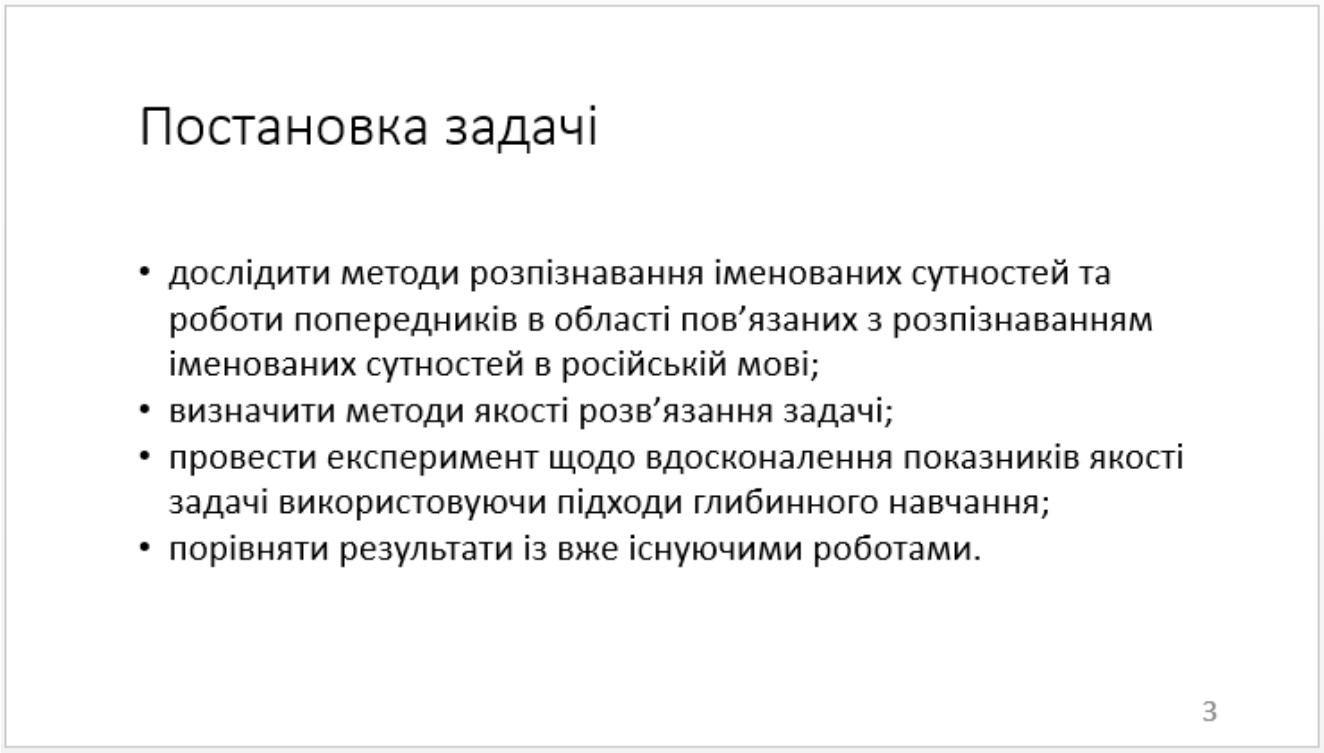


Рисунок А.3 – Слайд 3

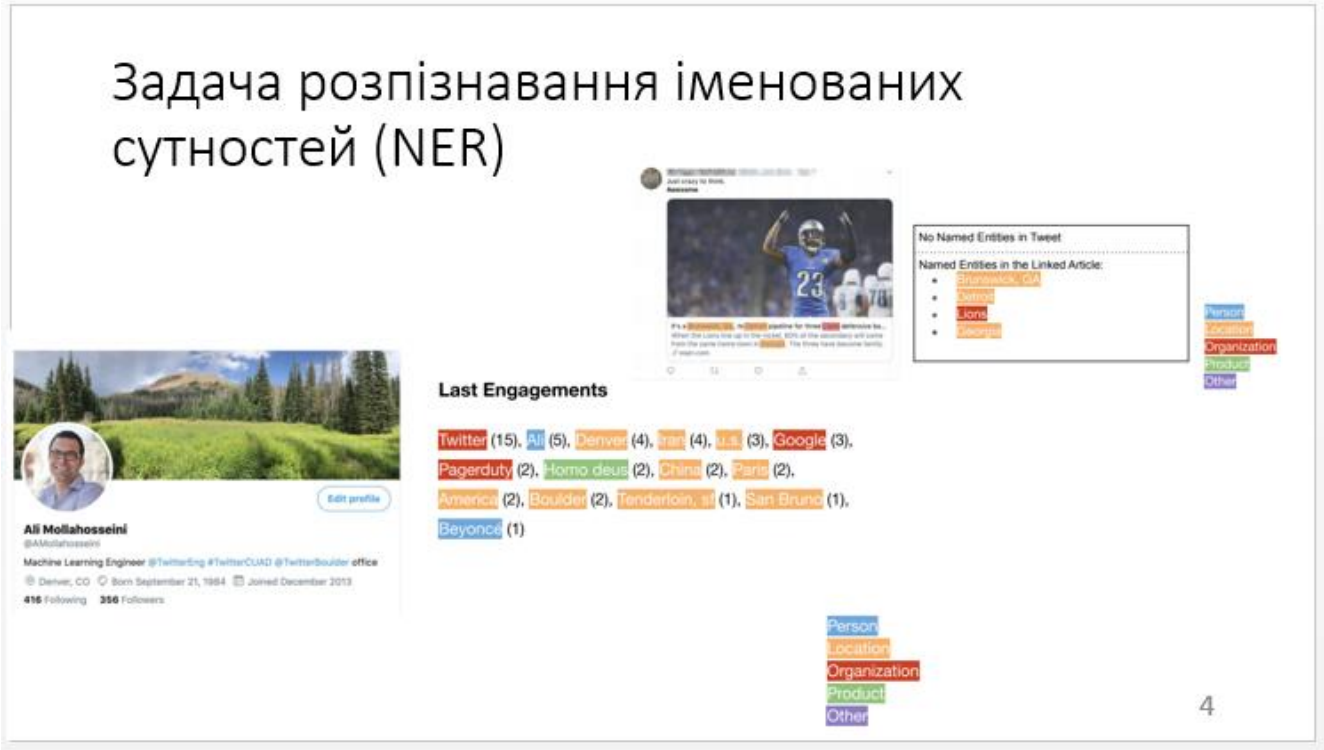


Рисунок А.4 – Слайд 4

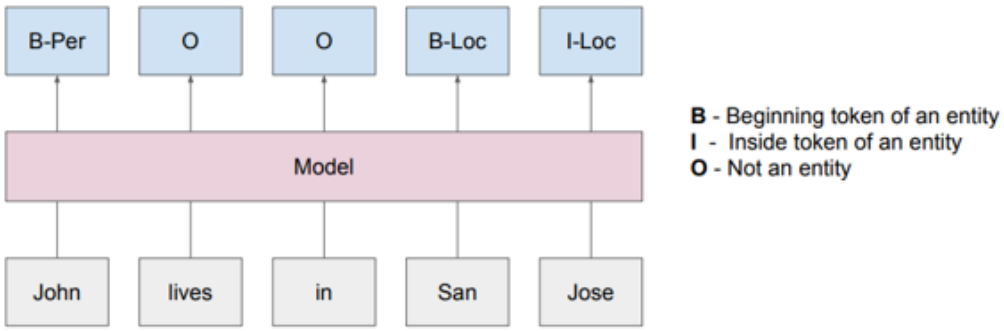
Складності при розв’язанні задачі NER

- анафора – визначення до якого іменника відноситься займенник;
- вільний порядок слів;
- неологізми - нові слова в мові;
- полісемія – безліч значень у одного слова (омоніми, омографи);

5

Рисунок А.5 – Слайд 5

Модель NER



6

Рисунок А.6 – Слайд 6

Існуючі методи розв’язання

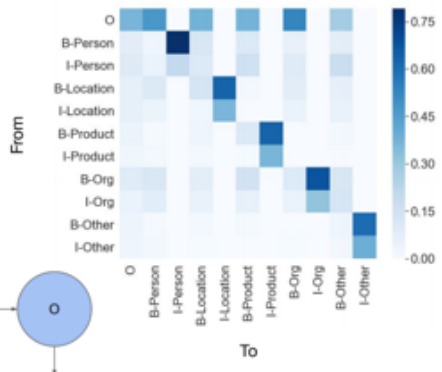
- підходи, що засновані на правилах;
- **методи машинного навчання**;
- **методи глибинного навчання**;

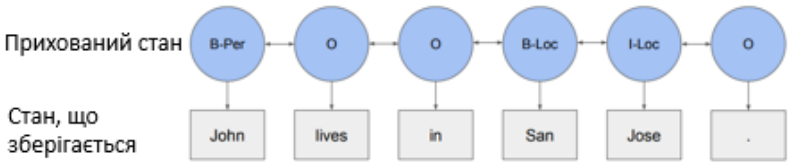
7

Рисунок А.7 – Слайд 7

Умовні випадкові поля (Conditional Random Fields, CRF)

- Дискримінаційний аналог прихованої моделі Маркова (HMM)
- Моделі локального контексту з матрицею переходів





8

Рисунок А.8 – Слайд 8

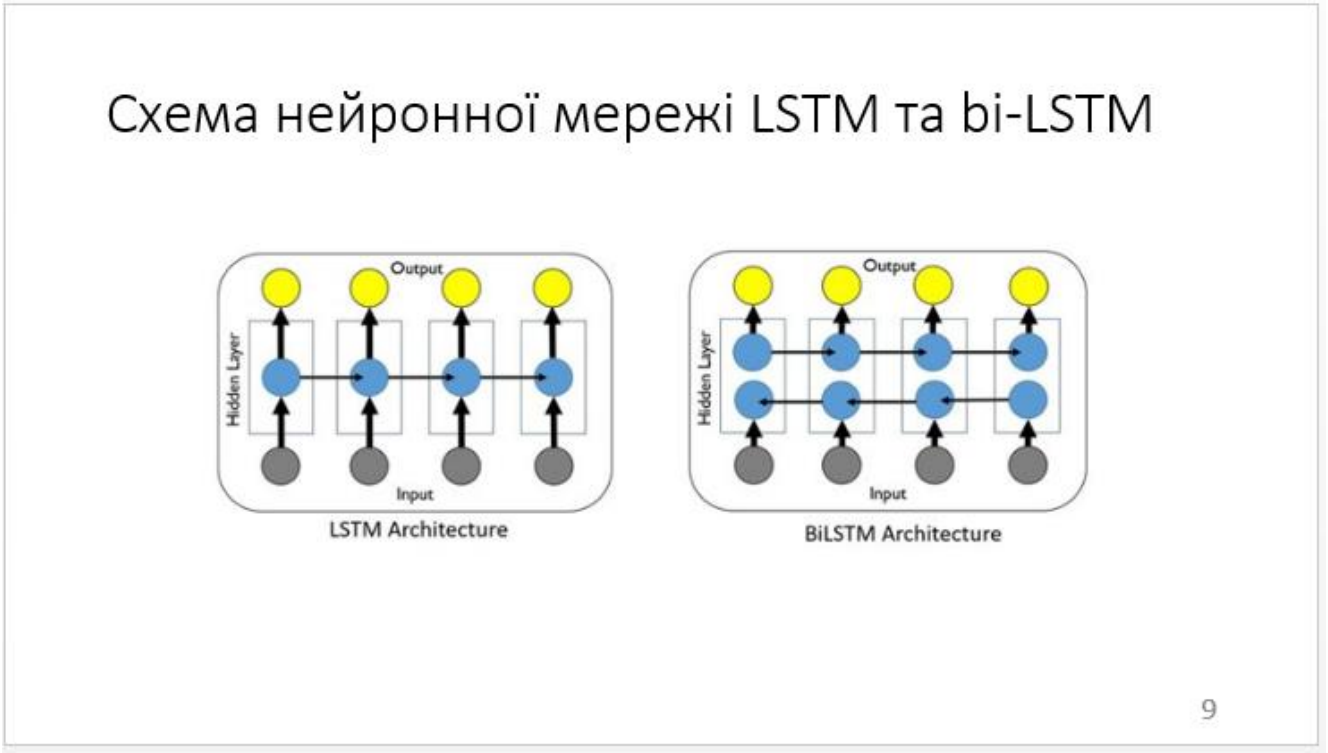


Рисунок А.9 – Слайд 9

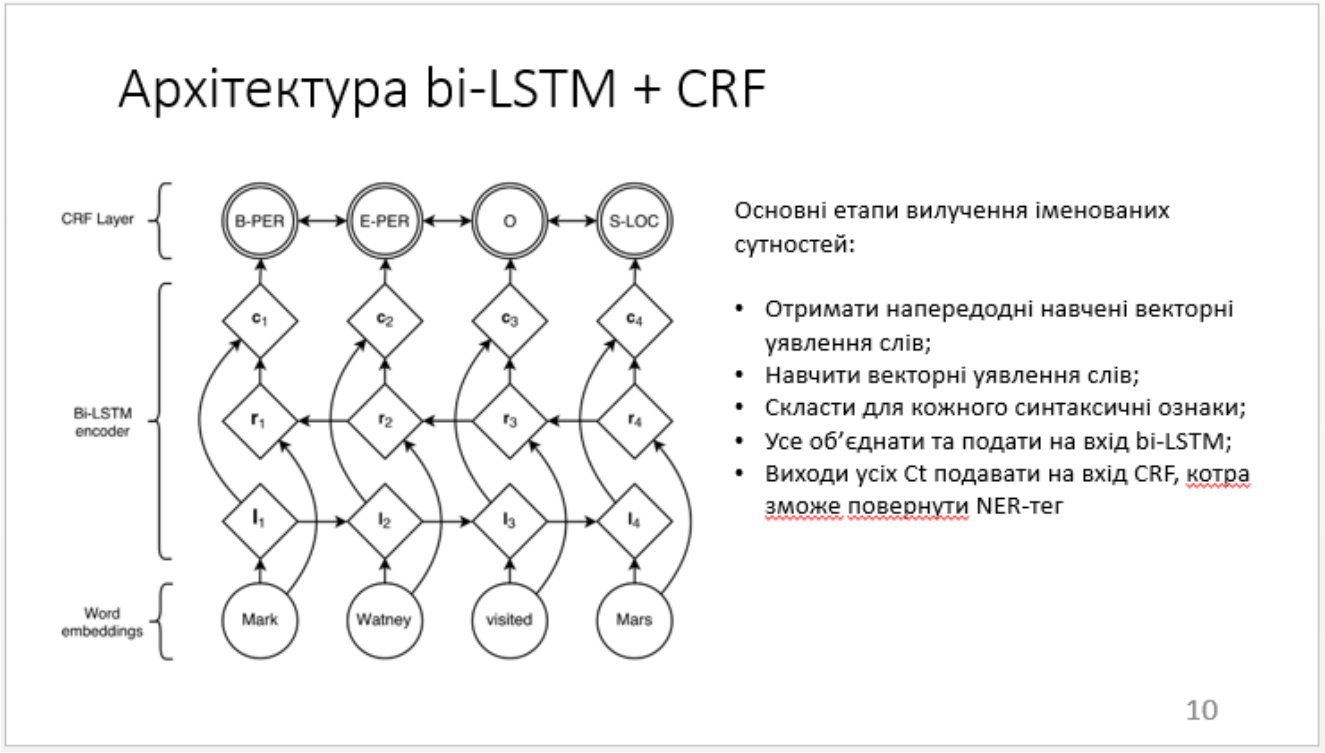
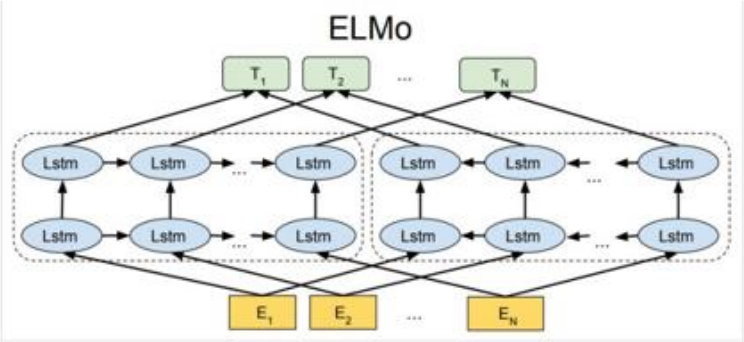


Рисунок А.10 – Слайд 10

Мовна модель Elmo



11

Рисунок А.11 – Слайд 11

Методи оцінки якості розв’язання задачі

- Recall;
- Precision;
- F1-score

12

Рисунок А.12 – Слайд 12

Методи оцінки якості розв’язання задачі

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F = 2\frac{Precision * Recall}{Precision + Recall}$$

TP - істинно-позитивні,
FN – хибно-негативні,
FP – хибно-позитивні

13

Рисунок А.13 – Слайд 13

Порівняння результатів

	Precision, %	Recall, %	F1-score, %
Wiki-based-approach	88.19	64.75	74.67
basic+dictionary + w2v features on SVM	82.57	74.08	78.10
SVM+w2v	-	-	82.88
Bi-LSTM + CRF + Lenta	83.80	80.84	82.10
NeuroNER + Highway char	80.59	80.72	80.66
Stanford Core NLP	71.01	64.19	67.13
RNN+MLP	72.90	76.60	74.71
Bi-LSTM-CRF	77.23	85.19	81.02
Bi-LSTM-CRF + ELMo	82.32	84.04	83.17
Bi-LSTM-CRF+ ELMo + fine tuning	83.19	85.41	84.29

14

Рисунок А.14 – Слайд 14

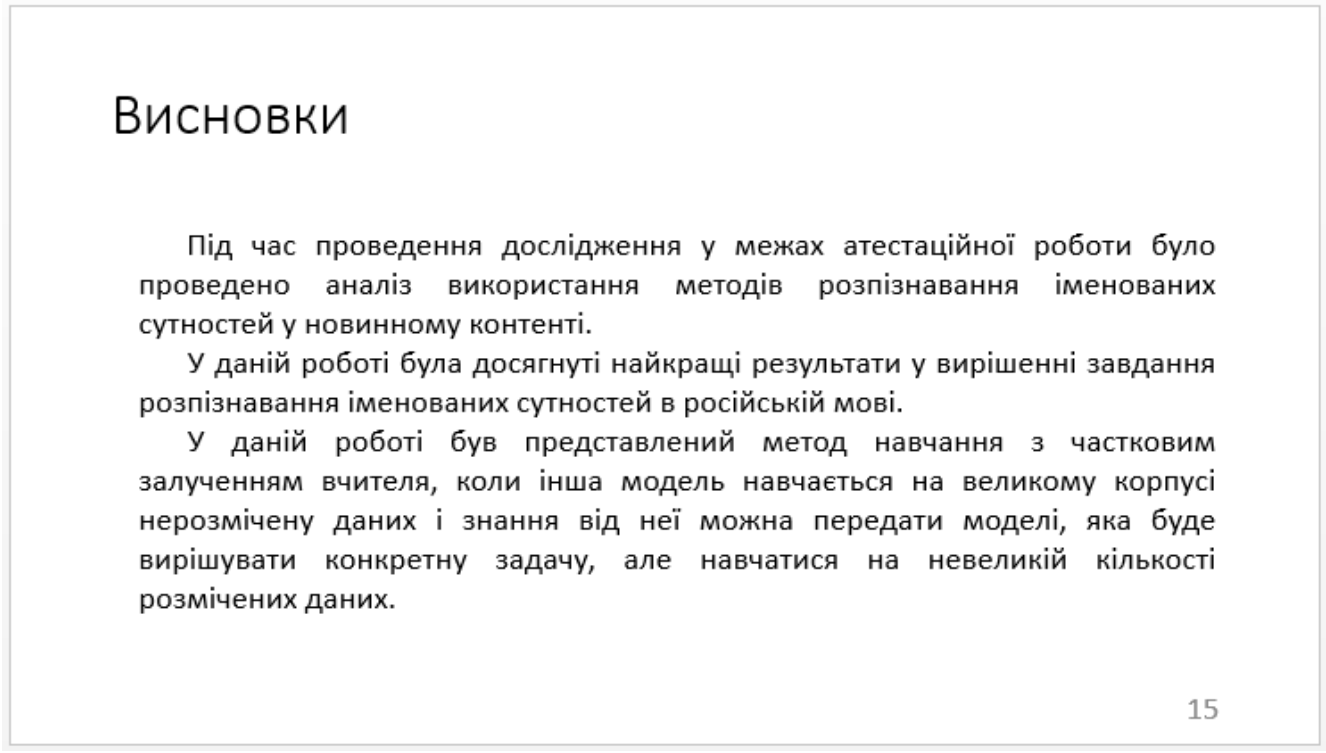


Рисунок А.15 – Слайд 15

Рисунок А.16 – Слайд 16

ДОДАТОК Б

Апробація результатів атестаційної роботи

SCI-CONF.COM.UA

TOPICAL ISSUES OF THE DEVELOPMENT OF MODERN SCIENCE

ABSTRACTS OF IX INTERNATIONAL SCIENTIFIC AND PRACTICAL CONFERENCE MAY 6-8, 2020

SOFIA 2020

УДК 004.89

ПРОБЛЕМИ РОЗПІЗНАВАННЯ ІМЕНОВАНИХ СУТНОСТЕЙ КОМП'ЮТЕРНИМИ СИСТЕМАМИ

Кожешкурт Ірина Вікторівна
студент
Харківский національный університет радіоелектроніки
Україна, м. Харків
Валенда Наталія Анатоліївна
к.т.н., доцент
Харківский національный університет радіоелектроніки
Україна, м. Харків

Анотація: Стаття присвячена огляду такої задачі обробки природньої мови як розпізнавання іменованих сутностей, а також огляд проблем які вникають при розпізнаванні тексту комп'ютерними системами та огляду сучасних методів їх вирішення.

Ключевые слова: обробка природної мови, розпізнавання іменованих сутностей, нейронні мережі, згорткові нейронні мережі.

Зі збільшення обсягів інформації доступної кожному користувачу глобальної мережі Інтернет виникає потреба в автоматизованому аналізі даних. Для розуміння проблеми дамо визначення текстової інформації з точки зору людини та комп'ютерної системи.

Для людини текст – це набір значимих фактів, з яких можна зробити певні висновки. Для комп'ютерної системи текст – це набір символів. Отже, основна проблема автоматизованого аналізу даних, це перетворення набору символів у набір значимих фактів, однаково зрозумілих людині та комп'ютерній системі, використовуючи які можна буде зробити певні висновки.

Обробка природної мови (англ. *Natural-language processing, NLP*) — загальний напрям інформатики, штучного інтелекту та математичної лінгвістики. Він вивчає проблеми комп'ютерного аналізу та синтезу природної мови. Стосовно штучного інтелекту аналіз означає розуміння мови, а синтез — генерацію розумного тексту. Розв'язок цих проблем буде означати створення зручнішої форми взаємодії комп'ютера та людини. [1, с.132]

Одним з найпопулярніших завдань обробки природної мови є розпізнавання іменованих сутностей в тексті.

Вирішення цієї задачі, по-перше, дає змогу знайти важливі для якогось завдання фрагменти тексту. Наприклад, можемо виділити тільки ті абзаци, де зустрічаються сутності якогось певного типу, а потім працювати тільки з ними.

В літературі найчастіше розглядають чотири типи сутностей: LOC, PER, ORG, MISC – місцевість, персона, організація та інше. [2, с.62]

На рисунку 1 зображено текст з розміченими сутностями.

Рис. 1 – Приклад вилучення іменованих сутностей із новинного контенту

Вирішення задачі розпізнавання іменованих сутностей - це крок до «розуміння тексту». Це може як мати самостійну цінність, так і допомогти краще вирішувати інші завдання NLP. Наприклад, побудову питально-відповідних систем.

Класичною складністю, яка виникає при вирішенні найрізноманітніших завдань NLP, є різного роду неоднозначності в мові. Наприклад, багатозначні слова й

омоніми. Є й окремий вид омонімії, що має безпосереднє відношення до задачі NER - одним і тим же словом можуть називатися зовсім різні сутності. Наприклад: *[Джек Лондон] (PER) народився в [Сан Франциско](LOC), [Каліфорнція](LOC), а не в [Лондоні](LOC).*

Різні входження слова «Лондон» відповідають різним типам іменованих сутностей – географічній назві та прізвищу (власній назві). Вирішення подібних ситуацій робить завдання виділення сутностей нетривіальним для вирішення простим алгоритмічним шляхом.

Для оцінки якості алгоритмів розпізнавання іменованих сутностей було запроваджено декілька критеріїв. Зазвичай до цих критеріїв входять повнота та точність. Ці критерії оцінки є широковживаними при оцінці алгоритмів класифікації.

Але проблема в тому, що при розпізнавані іменованих сутностей виникають часткові знаходження, які не можна віднести ні до хибних ні до вірних:

– сутність знайдено, але не всі токени сутності було вибрано;

– сутність знайдено, але було виділено також токени, які не відносяться до даної сутності;

– дві суміжні сутності однієї категорії виділено як одну;

– одну сутність знайдено як дві одного типу - сутність знайдено, проте класифіковано невірно.

Методи вирішення даної задачі можна поділити на 3 групи: основані на побудованих базах знань людиною, на застосуванні підходів машинного навчання та комбіновані.

Сучасні методи розв'язку задачі розпізнавання іменованих сутностей використовують рекурентні (RNN) та згорткові (CNN) нейронні мережі, а також умовні випадкові поля (Conditional Random Fields).[3, с.56]

Переваги методів основаних на використанні нейронних мереж:

– немає необхідності залучати знавців мови;

– дають високі показники при гарному наборі для тренування;

– створення набору для тренування достатньо мати носія мови.

Недоліки методів основаних на використанні нейронних мереж:

– для створення тренувальних даних потрібно багато часу;

– відсутність розуміння як конкретно нейронна мережа робить висновки;

– багато часу йде на вибір найбільш вдалої структури нейронної мережі для задачі.

Таким чином, вирішення задачі розпізнавання іменованих сутностей із використанням нейронних мереж є перспективною та потребує уваги для пошуку найбільш вдалої структури нейронної мережі для подолання проблем пов'язаних з особливостями мови.

СПИСОК ЛІТЕРАТУРИ

1. Bian, J., Gao, B., and Liu, T.-Y. (2014). Knowledge-powered deep learning for word embedding. In Machine Learning and Knowledge Discovery in Databases

2. Winograd, Terry (1971). Procedures as a Representation for Data in a Computer Program for Understanding Natural Language

3. Bethge, Matthias; Ecker, Alexander S.; Gatys, Leon A. (2015). A Neural Algorithm of Artistic Style

ДОДАТОК В
Відгук та рецензії

ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ РАДІОЕЛЕКТРОНІКИ

Факультет комп'ютерних наук

ВІДГУК

на атестаційну роботу магістра
Кожешкурт Ірини Вікторівни, ІПЗм-18-3,
спеціальність *121- Інженерія програмного забезпечення*
Освітньо-наукова програма *«Інженерія програмного забезпечення»*

Тема атестаційної роботи «Дослідження методів машинного навчання для розпізнавання іменованих сутностей у новинному контенті»

У атестаційній роботі магістра Кожешкурт І.В. розглянуто найпопулярніші методи машинного навчання для задачі розпізнавання іменованих сутностей.

Під час виконання роботи магістром були проаналізовані існуючі підходи до розв'язання задачі іменованих сутностей, представлений метод навчання з частковим залученням вчителя, коли інша модель навчається на великому корпусі нерозмічених даних і знання від неї можна передати моделі, яка буде вирішувати конкретну задачу, але навчатися на невеликій кількості розмічених даних. Для цього магістр опрацювала велику кількість літератури та веб-сторінок, що дозволило обґрунтовано сформулювати проблеми в цій галузі та запропонувати її рішення.

За час атестаційної роботи Кожешкурт І.В. продемонструвала самостійність, здатність до аналізу предметної галузі, спроможність до роботи з науково-технічною літературою, здібність використовувати сучасні засоби розробки.

Завдання виконувалися без відхилень від календарного плану. Роботу було виконано своєчасно та у відповідності до поставленої задачі. Пояснювальна записка виконана згідно вимогам.

Магістр гр. ІПЗм-18-3 Кожешкурт І.В готова до самостійної інженерної діяльності.

Атестаційну роботу можна подати до захисту в ЕК за спеціальністю 121 - «Інженерія програмного забезпечення», освітньо-науковою програмою Інженерія програмного забезпечення.

«18» 05. 2020р.

підпис

Керівник атестаційної роботи магістра
к.т.н, доц. Валенда Н.А.

Рецензія

на атестаційну роботу магістра
студента групи ІПЗм-18-3 Кожешкурт Ірини Вікторівни
спеціальність – 121- Інженерія програмного забезпечення
освітньо-наукова програма «Інженерія програмного забезпечення»

«Дослідження методів машинного навчання для розпізнавання іменованих сутностей у новинному контенті»
(Тема атестаційної роботи)

Структура атестаційної роботи: пояснювальна записка 59 сторінок; графічна частина 16 слайдів; програмне застосування (прикладна програма) 6 файлів загальним обсягом 2.0 Мбайт.

Представлена атестаційна робота магістра відповідає затвердженій темі та виконана відповідно до завдання.

Тема є актуальною за якою постійно проводяться нові дослідження.

Задачами дослідження були: аналіз методів машинного навчання для розпізнавання іменованих сутностей в тексті, проведення експерименту з ефективності їх застосування у текстах на російській мові, висновки щодо перспектив застосування отриманих результатів.

Обсяг роботи є достатнім та відповідає вимогам до випускних робіт магістрів. Розділи добре структуровані, змістовні. Надані усі необхідні додатки, що допомагають повною мірою оцінити виконану роботу. Цитування використаних джерел відповідає контексту. Джерела включають актуальні наукові публікації та патенти.

З урахуванням складності, робота відповідає вимогам до випускної атестаційної роботи магістра та має потенціал щодо впровадження результатів роботи на практиці у галузі обробки природньої мови при розв'язання задач.

Студентка Кожешкурт І.В. детально проаналізувала питання пов'язані з темою атестаційної роботи, розкрила деталі пов'язані зі специфікою методів машинного навчання та провела дослідження щодо застосування цих методів для задачі розпізнавання іменованих сутностей.

Результати роботи наочно і досить повно відображені в пояснювальній записці та на слайдах презентації. Пояснювальна записка написана грамотно, якість оформлення – висока, вимоги стандартів дотримані.

До недоліків атестаційної роботи слід віднести направленість дослідження лише на тексти російською мовою.

Студентка Кожешкурт І.В. провела глибокий аналіз спеціалізованої літератури та інтернет-ресурсів, опрацювала цю інформацію та обґрунтувала прийняті в роботі рішення.

Атестаційна робота магістранта групи ІПЗм-18-3 Кожешкурт І.В. відповідає вимогам до атестаційних робіт і заслуговує оцінки «відмінно – 90 В».

Магістрант гр. ІПЗм-18-3 Кожешкурт І.В. готова до самостійної інженерної діяльності.

Атестаційну роботу можна подати до захисту в ЕК за спеціальністю 121 – «Інженерія програмного забезпечення», освітньо-наукова програма «Інженерія програмного забезпечення».

Рецензент: к.т.н., доц., проф. каф. ШІ, ХНУРЕ Рябова Наталія Володимирівна

Рецензія

на атестаційну роботу магістра
студента групи ІПЗм-18-3 Кожешкурт Ірину Вікторівну
спеціальність – 121- Інженерія програмного забезпечення
освітньо-наукова програма «Інженерія програмного забезпечення»

«Дослідження методів машинного навчання для розпізнавання іменованих сутностей у новинному контенті»
(Тема атестаційної роботи)

Структура атестаційної роботи: пояснювальна записка 59 сторінок; графічна частина 16 слайдів; програмне застосування (прикладна програма) 6 файлів загальним обсягом 2.0 Мбайт.

Представлена атестаційна робота магістра відповідає затвердженій темі та виконана відповідно до завдання.

Тема є актуальною за якою постійно проводяться нові дослідження.

Обсяг роботи – достатній, відповідає вимогам до випускних робіт магістрів. Розділи добре структуровані, змістовні. Надані усі необхідні додатки, що допомагають повною мірою оцінити виконану роботу.

З урахуванням складності, робота відповідає вимогам до випускної атестаційної роботи магістра та має потенціал щодо впровадження результатів роботи на практиці у галузі обробки природньої мови.

Студентка Кожешкурт І.В. детально проаналізувала питання пов'язані з темою атестаційної роботи, розкрила деталі пов'язані зі специфікою методів машинного навчання та провела дослідження щодо застосування цих методів для задачі розпізнавання іменованих сутностей.

Дослідження, проведені в роботі є доцільними та можуть бути застосовані на практиці.

У роботі були проаналізовані існуючі підходи до розв'язання задачі іменованих сутностей, представлений метод навчання з частковим залученням вчителя, коли інша модель навчається на великому корпусі нерозмічених даних і знання від неї можна передати моделі, яка буде вирішувати конкретну задачу, але навчатися на невеликій кількості розмічених даних.

Недоліками роботи є те, що було досліджено не всі мовні моделі, які потенційно можуть давати також гарні результати.

Попри зазначений недолік, студентка провела достатньо комплексний аналіз та отримала вагомі результати, шо відображені у пояснювальній записці до атестаційної роботи.

Студентка Кожешкурт І.В. провела глибокий аналіз спеціалізованої літератури та інтернет-ресурсів, опрацювала цю інформацію та обґрунтувала прийняті в роботі рішення.

Атестаційна робота магістранта групи ІПЗм-18-3 Кожешкурт І.В. відповідає вимогам до атестаційних робіт і заслуговує оцінки «відмінно – 90 В».

Магістрант гр. ІПЗм-18-3 Кожешкурт І.В. готова до самостійної інженерної діяльності.

Атестаційну роботу можна подати до захисту в ЕК за спеціальністю 121 – «Інженерія програмного забезпечення», освітньо-наукова програма «Інженерія програмного забезпечення».

Рецензент: д.т.н., професор, завідувач кафедри інформаційних технологій проектування Національний аерокосмічний університет ім. М.Є. Жуковського «ХАІ» Дружинін Євген Анатолійович