Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук
(повна назва)

Кафедра Програмної інженерії
(повна назва)

АТЕСТАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти – другий (магістерський)

Дослідження методів машинного навчання для розпізнавання іменованих сутностей у новинному контенті

(тема)

Виконала: студентка 2 курсу, групи ІПЗм-18-3
Кожешкурт І.В.
(прізвище, ініціали)

спеціальності 121 – Інженерія програмного забезпечення
(код і повна назва спеціальності)

Освітньо-наукової програми
(тип програми)

Інженерія програмного забезпечення
(повна назва освітньої програми)

Керівник доц., к.т.н Валенда Н.А.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри, проф. _____ З.В.Дудар

2020 р.

ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ РАДІОЕЛЕКТРОНІКИ

Факультет Комп'ютерних наук
Кафедра Програмної інженерії
Рівень вищої освіти – другий (магістерський)

Спеціальність 121 – Інженерія програмного забезпечення
(код і повна назва)
Тип програми освітньо-наукова програма

Освітня програма Інженерія програмного забезпечення

ЗАТВЕРДЖУЮ:

Зав. кафедри _____

(підпис)

«_____»_____ 20 ___ р.

ЗАВДАННЯ

НА АТЕСТАЦІЙНУ РОБОТУ

студентові Кожешкурт Ірині Вікторівні
(прізвище, ім'я, по батькові)

1. Тема роботи Дослідження методів машинного навчання для розпізнавання іменованих сутностей у новинному контенті
затверджена наказом університету від "___"_____20___р. № _____

2. Термін подання студентом роботи до екзаменаційної комісії 20 червня 2020 р.

3. Вихідні дані до роботи *алгоритми застосування методів машинного навчання для розпізнавання іменованих сутностей, пояснювальна записка. Використовувати ОС Windows*

4. Перелік питань, що потрібно опрацювати в роботі *мета роботи, аналіз стану проблеми і постановка задачі, огляд методів для розпізнавання іменованих сутностей , методи глибинного машинного навчання, методи оцінки якості*

5 Консультанти розділів роботи

Найменування розділу	Консультант (посада, прізвище, ім’я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата
Спецчастина	доц., к.т.н. Валенда Н.А.		

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка *
1.	Аналіз предметної галузі	27 січня 2020 р.	виконано
2.	Огляд існуючих методів машинного навчання для розпізнавання іменованих сутностей	20 лютого 2020 р.	виконано
3.	Огляд методів глибинного навчання для розпізнавання іменованих сутностей	10 березня 2020 р.	виконано
4.	Підготовка пояснювальної записки	23 березня 2020 р.	виконано
5.	Спецчастина	24 квітня 2020 р.	виконано
6.	Підготовка презентації та доповіді	29 квітня 2020 р.	виконано
7.	Нормоконтроль, рецензування	10 травня 2020 р.	виконано
8.	Занесення диплома в електронний архів	15 травня 2020 р.	виконано
9.	Попередній захист	18 травня 2020 р.	виконано
10.	Допуск до захисту у зав. кафедри	20 травня 2020 р.	виконано

Дата видачі завдання _____ _____ 2020 р.

Студент _____
(підпис)

Керівник роботи _____ доц. Валенда Н.А.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ/ ABSTRACT

Пояснювальна записка до атестаційної роботи містить: 63 с., 10 рис., 1 таблиця, 22 джерел.

РОЗПІЗНАВАННЯ ІМЕНОВАНИХ СУТНОСТЕЙ, НОВИННИЙ КОНТЕНТ, МАШИННЕ НАВЧАННЯ, НЕЙРОННІ МЕРЕЖІ.

Метою роботи є дослідження методів машинного навчання для задачі розпізнавання іменованих сутностей.

У результаті роботи було розглянуто найпопулярніші методи машинного навчання, також було розглянуто методи що використовуються для вирішення задач пов'язаних з обробкою природної мови. Було побудовано та натреновано моделі, що можуть використовуватись для подальшого розпізнавання іменованих сутностей.

Було виявлено, що незважаючи не то, що розпізнавання іменованих сутностей не є новим у сфері комп'ютерної лінгвістики, проте результати розв'язання цієї задачі можна покращувати, особливо коли мова йде про датасети на різних мовах.

NAMED ENTITY RECOGNITION, NEWS CONTENT, MACHINE LEARNING, NEURAL NETWORKS.

The aim of the work is to study the methods of machine learning for the problem of recognizing named entities.

As a result, the most popular methods of machine learning were considered, as well as the methods used to solve problems related to natural language processing. Models have been built and trained that can be used to further identify named entities.

It has been found that while recognizing named entities is not new to computational linguistics, the results of this problem can be improved, especially when it comes to datasets in different languages.

ЗМІСТ

ВСТУП

В даний час відбувається експоненціальне зростання обсягів інформації в неструктурованою формі. Всесвітня мережа складається в основному з неструктурованих документів, з яких важко отримати корисну інформацію. Людських ресурсів для аналізу і обробки такого великого кількості інформації катастрофічно недостатньо. З огляду на все це, наявність алгоритмів і методів, здатних проводити аналіз інформації, структурувати і представляти її в зрозумілому людині вигляді, стає критичної завданням. Знання, що містяться неструктурованих документах, можуть бути доступні для машинної обробки і приведення їх у структуровану форму.

Сучасний розвиток штучного інтелекту дозволяє успішно вирішувати ряд завдань в області обробки природної мови без участі людини. Обробка природної мови (natural language processing, NLP) – це важлива область досліджень в комп'ютерних науках, що займаються аналізом документів на основі безлічі теорій і технологій. Важливою частиною обробки природної мови є отримання інформації. Витяг інформації з тексту полягає в добуванні об'єктів, зв'язку між цими об'єктами і в кінцевому підсумку в добуванні необхідних фактів. Це процес можна описати як розуміння тексту без участі людини.

Зараз активно розвиваються діалогові системи, в яких розуміння тексту є однією з головних задач. Штучний інтелект в області обробки природної мови може допомагати операторам технічної підтримки знаходити швидше правильну відповідь або спілкуватися з людьми на початковій стадії діалогу. Також на новинних порталах важливо тримати все статті в структурованому вигляді, штучний інтелект може сам класифікувати документ на потрібну тему або запропонувати ряд відповідних тим на вибір людині. У соціальних мережах і інших джерелах, що містять розважальний контент, користувачеві постійно треба рекомендувати щось нове, а іноді і те, про що він сам не здогадується. Штучний інтелект може допомогти людині у вивченні іноземних мов, в сортуванні листів на

поштою, відповідати на деякі листи. Голосові команди зараз впроваджуються всюди, а це теж лежить в області обробки природної мови. В кінцевому підсумку хочеться прийти до того, щоб штучний інтелект розумів природна мова так само добре, як і людина. Міг застосовувати знання світу при їх аналізі.

Для створення всього вищеописаного майже завжди спершу необхідно вирішити задачу вилучення об'єктів з тексту або, іншими словами, розпізнати іменовані сутності в тексті (named entity recognition, NER).

Метою даної роботи є аналіз, порівняння та покращення точності існуючих методів та підходів до розпізнавання іменованих сутностей в неструктурованому тексті на російській мові.

Задачами дослідження є проведення експериментів спрямованих на виявлення методів, що мають найкращі показники на наборі даних новинного контенту на російській мові.

Об’єктом дослідження є іменовані сутності в тексті на російській мові

Предметом дослідження є методи машинного навчання для розпізнавання іменованих сутностей.

Елементом наукової новизни магістерської роботи є систематизація попередніх знань щодо розв’язання задачі іменованих сутностей, а також дослідження шляхів покращення точності при вирішення даної проблеми.

Результати проведеного дослідження можуть бути використані для розв’язку різних прикладних задач.

Частина роботи, що пов’язана із дослідженням проблем, що виникають при розв’язанні цієї проблеми, а також їх вплив на результати опублікована у збірнику тез з IX Международної науково-практичної конференції у м. Софія, Болгарія. Тези доповіді наведено в додатку Б.

1 АНАЛІЗ СТАНУ ПРОБЛЕМИ

1.1 Обробка природньої мови

Історія області обробки природної мови почалася ще в 1950-х роках. Однак, аж до 1980-х років більшість систем обробки природної мови були реалізовані на наборах складних рукописних правил. З впровадженням методів машинного навчання відбулася революція в цій області. Одним з ранніх методів використовувалися дерева рішення [1], які були альтернативою рукописним правилам.

Також стали популярні статистичні моделі, такі як прихована марківська модель [2]. Такі моделі були більш надійні на невідомих даних і на даних, які містили помилки (як це часто зустрічається в реальності) [21].

Зараз дуже популярні рішення, засновані на навчанні без учителя або з частковим залученням вчителя [3]. Такі алгоритми можуть витягати знання з немаркованих текстів, що дуже важливо, якщо врахувати зростання обсягу інформації. Глибокі нейронні мережі стали невід'ємною частиною обробки природної мови та дозволяють досягти найкращих результатів.

Область обробки природної мови вирішує безліч завдань:

- машинний переклад – це область обчислювальної лінгвістки, що досліджує застосування програмного забезпечення для перекладу тексту з одної мови на іншу. У простому наближенні це проста заміна слів однієї мови на іншу;
- питально-відповідні системи – системи, які можуть відповідати на питання. Для пошуку відповіді, наприклад, використовується Вікіпедія. За вхідного питання одна нейронна мережа звужує список статей, в яких шукати, а інша вже робить детальний пошук відповіді;
- розпізнавання мови – це область обчислювальної лінгвістики, яка вивчає застосування технологій для розпізнавання розмовної мови і його перекладу в текст;

- класифікація текстів – визначення тем документів. Важливе завдання в специфічних доменних областях;
- витяг фактів – процес аналізу документів на основі великої кількості теорій і технологій. Витяг об'єктів, зв'язок між цими об'єктами і в кінцевому підсумку добуванні необхідних фактів;
- діалогові системи – це комп'ютерні системи, призначені для спілкування з людиною. Ці системи складаються з великої кількості компонентів: розпізнавання мови, жестів, розуміння мови (Розуміння мови вже сама по собі комплексна задача), видача заготовленої або згенерованої відповіді, переклад на розмовну мова або на мову жестів;
- аналіз тональності тексту – визначення емоційного забарвлення, настрою тексту [22];
- суммарізація текстів – процес скорочення текстів, вилучення тільки важливої інформації;
- topic modeling - визначення переліку тем з великої колекції документів. Кожна тема представляється розподілом слів в ній;
- розв’язок кореферентності – визначення зв'язків між об'єктами і компонентами висловлювання;
- розв’язок лексичної багатозначності – визначення сенсу слова в залежності від контексту;
- визначення частин мови;
- синтаксичний розбір;
- приведення слова в початкову форму.

Задачі можна розділити на декілька категорій: морфологічні, синтаксичні, семантичні та прагматичні.

На рисунку 1.1 представлена піраміда задач: чим вище рівень, тим більше складність задач.

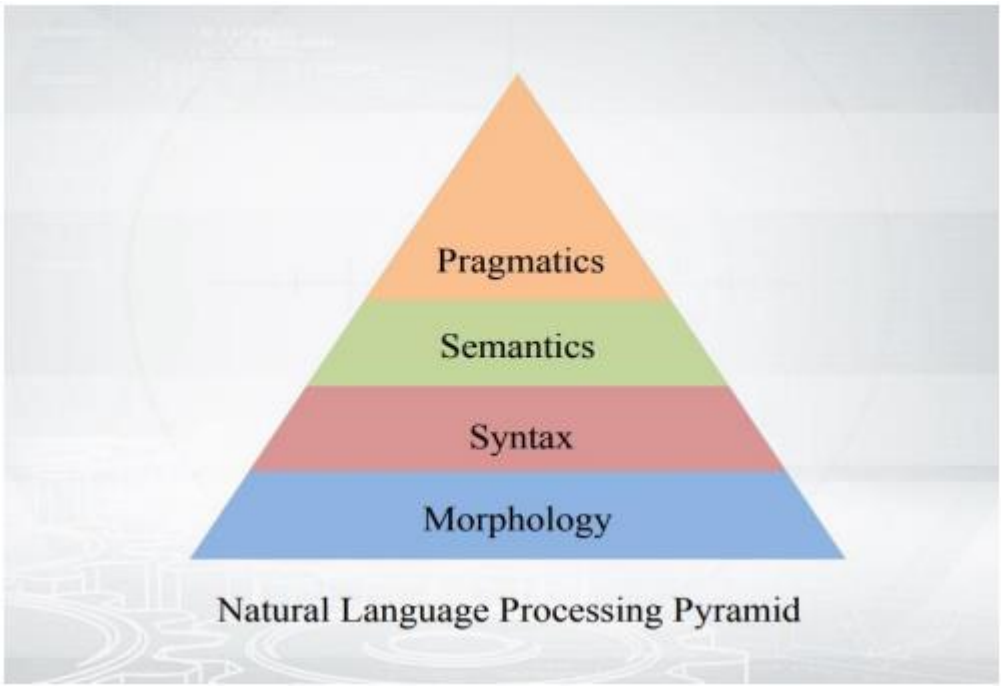

Рисунок 1.1 – Піраміда задач в обробці природньої мови

Перші три категорії задач відноситься до граматичної мови – наука, вивчає закономірності побудованих вірних та осмислених пропозицій та текстів. Прагматика вивчає відношення знаків до суб’єктів, які їх випромінюють і інтерпретують.

Складнощі, що виникають при обробці природної мови:

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

В даній роботі ми більш детально розглянемо задачу розпізнавання іменованих сутностей та методи її розв’язання.

1.1.1 Історія та суть задачі розпізнавання іменованих сутностей

На шостий конференції Message Understanding (MUC-6) в 1996 році завдання фокусувалися в області добування інформації. В процесі постановки завдань виявилася окреме завдання в добуванні об'єктів з документів. Для визначення об'єкта ввели термін «іменована сутність», а задачу назвали розпізнавання іменованих сутностей (named entity recognition, NER), так її в області обробки природної мови і називають дотепер. Тепер потрібно було вміти розпізнавати в тексті різні сутності: імена, організації, числові вирази, розташування, дати, час і багато іншого.

На рисунку 1.2 зображено текст з розміченими сутностями.

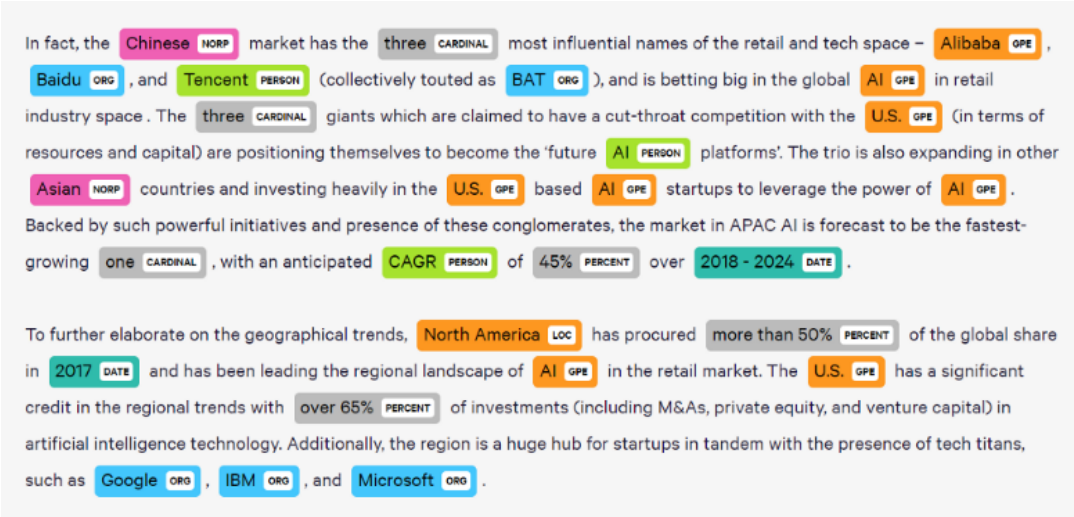

Рисунок 1.2 – Приклад розпізнавання іменованих сутностей у тексті

Розв’язання задачі розпізнавання іменованих сутностей може бути використана для таких потреб:

– класифікація новинного контенту. Новини та видавництва щодня генерують велику кількість інтернет-контенту, а правильне управління ними дуже важливо, щоб максимально використовувати кожну статтю. Розпізнавання іменованих може автоматично сканувати цілі статті та розкривати основних людей, організацій та місця, в яких обговорюється.

Знання відповідних тегів для кожної статті допомагає автоматично класифікувати статті за визначеними ієрархіями та дозволяє легко розкривати вміст;

– ефективні алгоритми пошуку. Якщо для кожного пошукового запиту алгоритм закінчує пошук усіх слів у мільйонах статей, процес займе багато часу. Натомість, якщо розпізнавання названої сутності може бути запущено один раз у всіх статтях, а відповідні об'єкти (теги), пов'язані з кожною з цих статей, зберігаються окремо, це може значно прискорити процес пошуку. При такому підході пошуковий термін буде узгоджений лише з невеликим списком сутностей, обговорюваних у кожній статті, що призводить до швидшого виконання пошуку;
– посилення контекстних рекомендацій. Один з найважливіших випадків використання розпізнавання іменованих сутностей включає автоматизацію процесу рекомендацій. Системи рекомендацій домінують у тому, як ми відкриваємо новий вміст та ідеї в сучасному світі. Приклад Netflix показує, що розробка ефективної системи рекомендацій може творити чудеса для статей медіа-компанії, роблячи їх платформи більш привабливими та захоплюючими подіями;
– підтримка клієнтів. Існує ряд способів зробити процес зворотного зв’язку з клієнтами плавним, і одним із них може стати розпізнавання іменованих сутностей. Якщо ви працюєте з відділом підтримки покупців електронного магазину з кількома філіями по всьому світу, ви переглядаєте ряд згадок у відгуках своїх клієнтів.
– класифікація наукових праць. Інтернет-журнал або веб-сайт видань містить мільйони наукових робіт та наукових статей. На одну тему може бути сотні робіт з невеликими відмінностями. Можна вдало класифікувати ці роботи за рахунок розпізнавання іменованих сутностей.

В даній роботі буде досліджено розв’язок задачі розпізнавання іменованих сутностей на приклади класифікації новинного контенту.

1.2 Огляд традиційних методи вирішення задачі розпізнавання іменованих сутностей

Традиційні підходи до розпізнавання іменованих сутностей класифікуються на три основні потоки: на основі правил, навчання без вчителя, навчання з вчителем.

Останнім часом домінуючими стають моделі розпізнавання іменованих сутностей, що засновані на методах глибинного навчання (Deep Learning).

У порівнянні з підходами, що базуються на ознаках, методи глибинного навчання можуть виявляти приховані ознаки автоматично.

Глибинне навчання – це галузь машинного навчання, яке складається із декількох шарів обробки для вивчення уявлень даних із кількома рівнями абстракції.[4] Типові шари – це штучні нейронні мережі із прямим та зворотнім напрямками.

1.2.1 Підходи, засновані на правилах

Даний підхід до пошуку іменованих сутностей в неструктурованому тексті полягає в ручному аналізі доменної області та визначенні шаблонів на основі яких буде відбуватись ідентифікація іменованих сутностей. Системи NER, що засновані на правилах, мають більшу точність, але ціною меншого запам’ятовування. Інколи такі системи залучають до роботи досвідчених лінгвістів. Кім запропонував використовувати підхід мовлення Brill для введення мови.[5] Це система автоматично генерує правила, засновані розмічування частин мови Брілла.

У біомедичній галузі запропоновано ProMiner, який використовує попередньо оброблений словник синонімів для виявлення в біомедичному тексті.

Також було запропоновано підхід, заснований на словнику для NER в галузі електронного здоров'я записи. Експериментальні результати показують, що такий підхід покращує повноту, але має обмежений вплив на точність.

Деякі інші відомі системи NER на основі правил включають LaSIE-II, NetOwl, Facile, SAR, системи FASTUS та LTG. Ці системи є в основному засновані на рукотворних смислових та синтаксичних правилах ідентифікації суб'єктів. Системи на основі правил працюють дуже добре коли лексикон обмежений. На рисунку 1.3 зображено архітектуру системи заснованої на правилах.

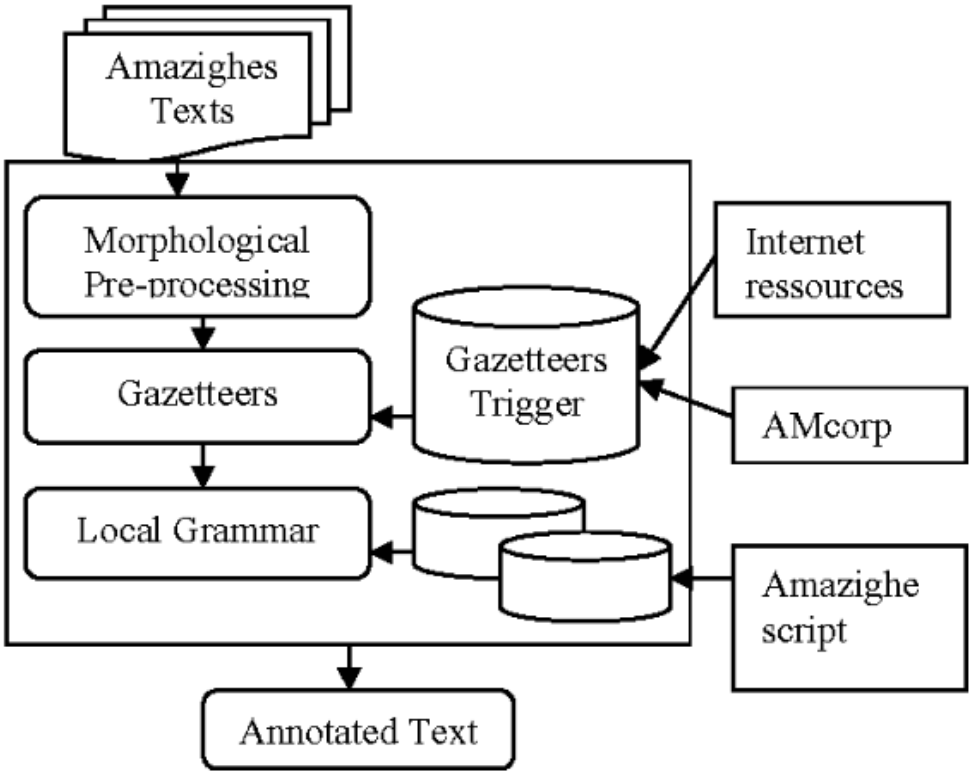

Рисунок 1.3 – Архітектура систем, заснованих на правилах

Так як системи, засновані на правилах, можуть бути використані тільки певній доменній області, потрібен значний час на перенавчання такої моделі.

1.2.2 Методи машинного навчання

Головною проблемою використання підходів заснованих на правилах є неможливість використовувати однакові правила в різних доменах. Щоб покращити такі підходи було винайдено підходи засновані на даних. Навчання з вчителем вимагає великої кількості анотованих даних, які складаються як з негативних , так і з позитивних прикладів іменованих сутностей. Модель навчається ознакам із анотованих корпусів для того, щоб створювати правила автоматично, які співпадають із конкретних типів іменованих сутностей. Популярними статистичними моделями є приховані марковські моделі, дерева рішень, методи опорних векторів, умовні випадкові поля.

Під час часткового навчання використовуються анотовані та немарковані дані. Починається все із використання деяких немаркованих даних під наглядом для «завантаження» процесу навчання. Наприклад, Брін (1988) вперше використав лексичні ознаки для побудови регулярних виразів, як основного правила для початку пошуку пар авторів книги у мережі Інтернет. Як тільки пара знайдена її можна буде також знайти в іншому форматі на тій же веб-сторінці. Наприклад, з формату «Марк Твен, Гекльберрі Фін ", може бути знайдений новий формат" Гекльберрі Фін, Марк Твен ". Тоді система може витягти нові правила на основі нового формату, щоб знайти більше пар. Рілофф, Джонс та ін. (1999) запропонували метод взаємного завантаження, який починається подаючи підготовлені приклади сутності певного типу до системи, а потім використовує контекст, знайдений навколо сутностей для створення нових моделей для пошуку нових сутностей.

Навчання без нагляду вимагає лише немаркованих даних, та основний підхід це кластеризація.

1.3 Методи глибинного навчання

В даний час дуже популярними є рішення, засновані на використанні нейронних мереж. Відкритим стоїть питання вибору ознак для нейронної мережі: ручні правила, морфологічні, синтаксичні та семантичні. Невід'ємною частиною завдань обробки природньої мови стало використання векторного уявлення слів, отриманого з різних моделей мови.

Як відомо, нейронна мережа вміє добре автоматично отримувати репрезентативні ознаки. Також рекурентна нейронна мережа кодує вхідну послідовність слів в контекстно-чутливе уявлення. Однак, для якісного представлення нейронної мережі потрібна велика навчальна вибірка. Для вирішення задачі розпізнавання іменованих сутностей сучасна архітектура нейронної мережі складається зазвичай з декількох двонаправлених рекурентних шарів [4]. Далі ознаки з останнього шару нейронної мережі можна подати в інший класичний алгоритм. У прикладі нижче представлений шар умовних випадкових полів (CRF Layer), в якості ознак до якого подаються ваги з виходу останнього шару двонаправленної нейронної мережі (див. рис. 1.4).

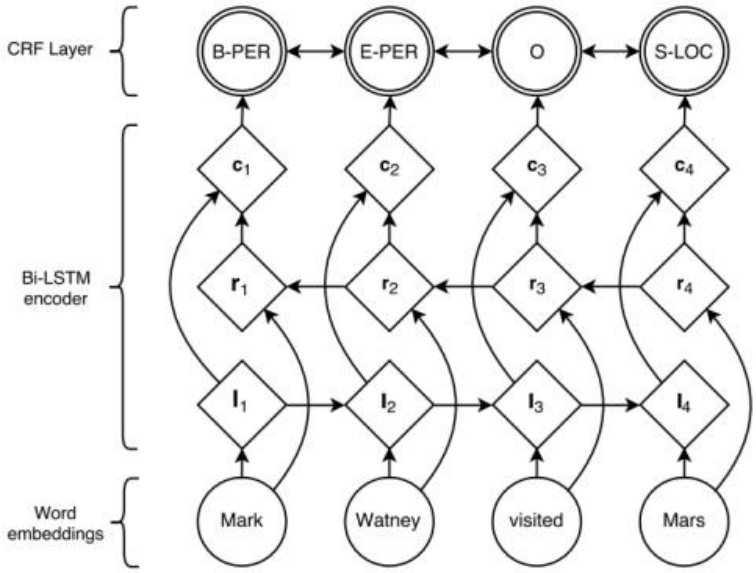

Рисунок 1.4 – Приклад системи нейронної мережі

Дуже популярні гібридні системи, які можуть поєднувати в собі всі типи методів для вирішення задачі розпізнавання іменованих сутностей.

1.4 Постановка задачі

В даній роботі необхідно буде дослідити існуючі методи розпізнавання іменованих сутностей та роботи попередників в області задачі пов’язаних з розпізнаванням іменованих сутностей в текстах на російській мові. Також необхідно визначити методи визначення якості розв’язання задачі. Провести експеримент щодо вдосконалення показників якості задачі використовуючи підходи глибинного навчання. Порівняти результати з вже існуючими на даний момент роботами на основі порівняння точності результатів. Дослідження буде проведено на FactRuEval корпусі даних

Виходячи з результатів, зробити висновок про подальші перспективи розвитку роботи.

2 ОПИС ПРОВЕДЕНИХ ТЕОРИТИЧНИХ ДОСЛІДЖЕНЬ

У цьому розділі буде дано опис всіх алгоритмів і архітектур нейронних мереж, які будуть використовуватися в практичних експериментах. Будуть описані різні векторні уявлення слів, які несуть в собі необхідні знання природної мови. Також будуть описані підходи навчання з частковим залученням вчителя і техніка тонкого налаштування, яке необхідне для досягнення високих результатів у вирішенні специфічних доменних завдань.

2.1 LSTM рекурентна нейронна мережа

Рекурентні нейронні мережі використовуються для вирішення великого числа завдань, включаючи завдання обробки природної мови, через їх здатність враховувати попередню інформацію з послідовності для розрахунку поточного виходу. Однак було виявлено [6], що, незважаючи на теоретичну можливість вивчати довгострокову залежність, на практиці моделі рекурентних нейронних мереж не виконують очікуваного і страждають від проблем з градієнтним спуском. З цієї причини була розроблена спеціальна архітектура нейронної мережі під назвою Long Short-Term Memory (LSTM) для вирішення проблеми загасання градієнта [7].

LSTM складається з наступних частин (див. рис .2.1):

- вхід мережі (input);
- вихід мережі (output);
- пам'ять або стан мережі (memory cell);
- блок очищення пам'яті (forget gate);
- блок оновлення пам'яті (input gate);
- блок видачі результату (output gate);

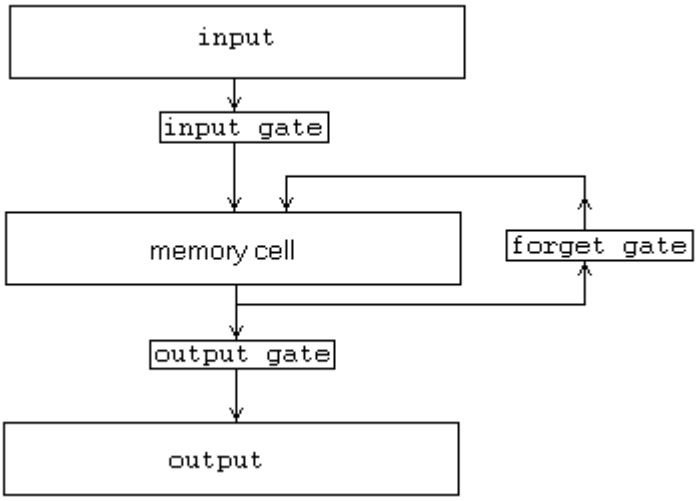

Рисунок 2.1 - Схема нейронної мережі LSTM

Формули компонентів представлені на формулі 2.1.

$$
\begin{aligned}
i_t &= \sigma(W_{ix}x_t + W_{ih}h_{t-1} + b_i),\\
f_t &= \sigma(W_{fx}x_t + W_{fh}h_{t-1} + b_{if}),\\
c_n &= g(W_{cx}x_t + W_{ct}h_{t-1} + b_c),\\
c_t &= f_t \circ c_{t-1} + i_t \circ c_n,\\
h_t &= o_t \circ g(c_t),\\
o_t &= \sigma(W_{ox}x_t + W_{oh}h_{t-1} + b_o),
\end{aligned}
\tag{2.1}
$$

де i, f, o, c – компоненти блоку пам’яті, σ – сігмоїдна функція, W – матриця вагів, g – гіперболічний тангенс, b – баєсовські вектори.

2.2 Двонаправлена LSTM

Правильне розпізнавання іменованого об'єкта в реченні залежить від контексту слова. Як попередні, так і наступні слова мають значення для прогнозування певної іменованої сутності. Двонаправлені рекурентні нейронні мережі (bi-LSTM) [8] були розроблені для кодування кожного елемента в послідовності з урахуванням лівого і правого контекстів, що робить їх одним з найкращих варіантів для вирішення завдання розпізнавання іменованих сутностей. Двонаправлений розрахунок моделі складається з двох етапів: прямий шар обчислює уявлення лівого контексту, і зворотний шар обчислює уявлення правого контексту. Виходи цих кроків потім об'єднуються для отримання повного уявлення елемента вхідної послідовності. Було показано, що bi-LSTM корисні в багатьох завданнях природної мови, таких як машинний переклад, питально-відповідні системи і особливо в розпізнаванні іменованих сутностей.

2.3 Умовні випадкові поля

Умовне випадкове поле (CRF) – імовірнісна модель для структурованого передбачення, яка успішно застосовується в різних областях, таких як комп'ютерне зір, біоінформатика, обробка природної мови. CRF може використовуватися незалежно для вирішення завдання розпізнавання іменованих сутностей [9].

CRF модель навчається для того, щоб передбачити вектор $y = \{y_{0,}\, y_1, \dots y_T\}$ тегів для даного речення $x = \{x_{0,} x_{1,} \dots x_T\}$. Для цього розраховується умовна ймовірність за формулою 2.2.

$$p(y|x) = \frac{e^{score(x,y)}}{\sum_{y'} e^{Score(x,y')}},$$
$$Score(x,y) = \sum_{i=0}^{T} A_{y_i,y_{i+1}} + \sum_{i=1}^{T} P_i, y_i \quad (2.2)$$

де $A_{y_i,y_{i+1}}$ означає імовірність, яка представляє собою оцінку від тега i до тега j, P_i, y_i – імовірність переходу, яка представляю собою j-го тегу i слова.

На етапі навчання логарифмічна ймовірність правильної послідовності міток $log\ (p\ (y \mid x))$ максимізується.

2.4 Поєднана bi-LSTM та CRF модель

Російська мова – це морфологічно і граматично багата мова. Таким чином, можна припустити, що поєднання CRF-моделі з вихідними вагами нейронної мережі bi-LSTM має підвищити точність виділених іменованих сутностей. Для кожного вхідного пропозиції мережу bi-LSTM видає послідовність оцінок, які представляють ймовірність кожного тега для кожного слова.

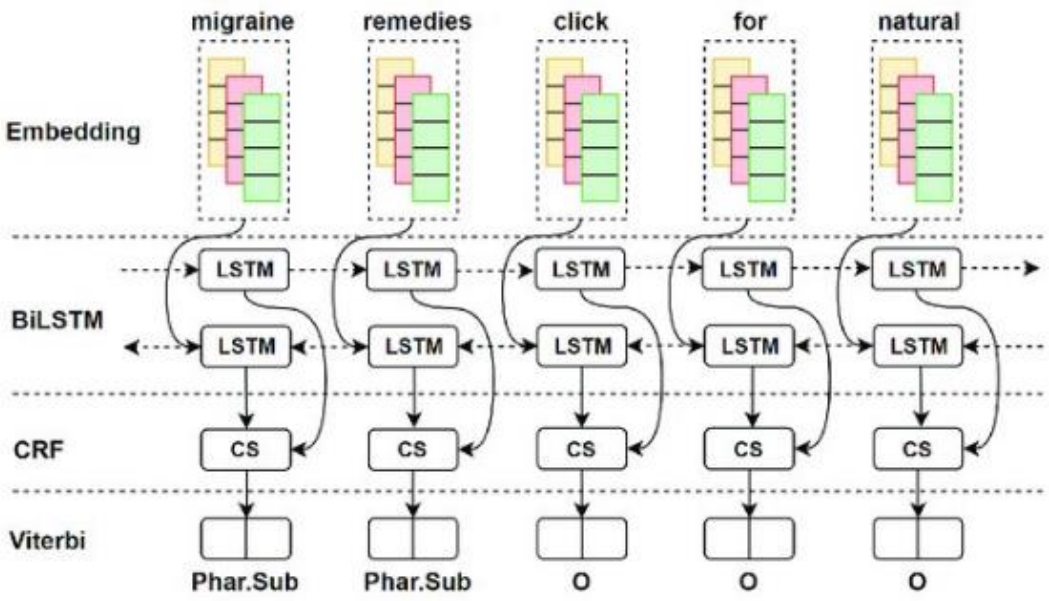

Рисунок 2.2 - Архітектура bi-LSTM+CRF

На рисунку 2.2 зображена архітектура bi-LSTM + CRF. Для підвищення точності прогнозів CRF шар [9] навчається з обмеженнями, залежними від порядку тегів.

2.5 Векторне уявлення слів

Як вже зазначалося вище, стандартним компонентом нейронної мережі для вирішення завдань обробки природної мови є векторні уявлення слів, що проходять попереднє навчання. В даний час дуже популярні дистрибутивні семантичні моделі, які дають єдине контекстно-незалежне уявлення слова [10].

В роботі досліджені глибокі контекстно-залежні уявлення слів, взяті з навченої двобічної мовної моделі і їх комбінація з дистрибутивними уявленнями. Також для підвищення точності прогнозів проводиться тонке налаштування ваг (fine tuning) двонаправдленої мовної моделі на текстах, які використовуються для

навчання моделі розпізнавання іменованих сутностей. При тонкому налаштуванні ваги в двонаправленій мовної моделі нам все також потрібні нерозмічену дані.

Однак, тонке налаштування ваг слід проводити акуратно, щоб нейронна модель не втратила знання, виділені при первісному навчанні. Описану мовну модель можна побачити на рисунку 2.3.

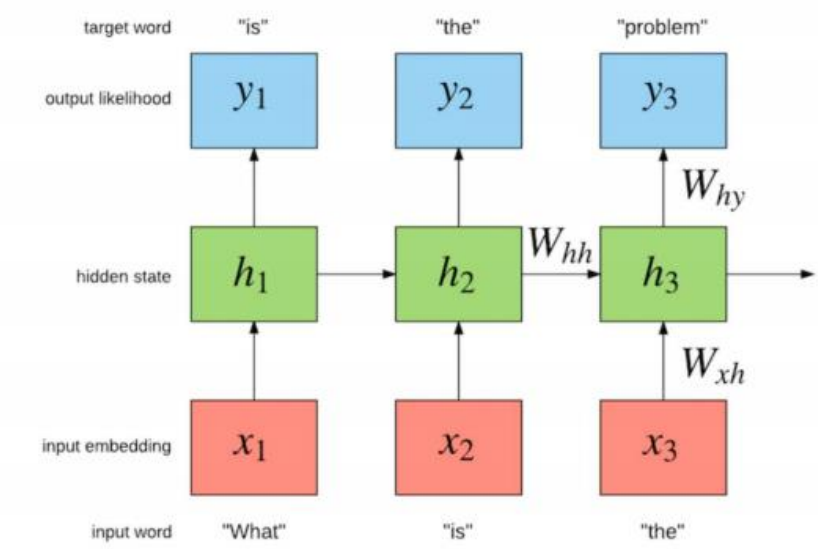

Рисунок 2.3 – Мовна модель, що представлена рекурентною нейронною мережею

Далі будуть наведені двонаправлені мовні імоделі з різними налаштуваннями.

2.5.1 Двонаправлена мовна модель

Мовна модель по даній послідовності слів передбачає, яке слово буде наступним.

Дана послідовність, яка містить в собі N токенів $(t_1, t_2, \dots, t_N)$, пряма мовна модель обчислює ймовірність послідовності шляхом моделювання ймовірності

токена t_k на основі попередніх токенів $(t_1, t_2, \dots, t_{k-1})$ та обчислюється за формулою 2.3

$$p(t_1,\ t_2, \dots, t_N) = \prod_{k=1}^{N} p(t_k|t_1, t2, \dots, t_{k-1}), \tag{2.3}$$

де $(t_1, t_2, \dots,\ t_N)$ – послідовність токенів мовної моделі

Сучасні мовні моделі обчислюють контекстно-незалежне уявлення токена x_k^{LM} (через буквене уявлення або через уявлення згорткової нейронної мережі), а потім передають його в пряму рекурентну мережу LSTM з L шарами. На кожній позиції k, кожен шар LSTM дає на виході контекстно-залежне уявлення $h_{k,j}^{\overrightarrow{LM}}$, де $j = 1, \dots, L$.

Вихід з верхнього LSTM шару, $h_{k,j}^{\overrightarrow{LM}}$, використовується для передбачення наступного токену t_{k-1} разом із шаром Softmax (узагальнення логістичної функції для багатовимірного випадку). Зворотня мовна модель схожа на пряму млвну модель, відміна лише в тому, що вона обробляє послідовність токенів в зворотному напрямку, передбачаючи попередній токен за даними наступних токенів. Це може бути реалізовано аналогічно прямої мовної моделі. Для кожного зворотного LSTM шару j модель виробляє контекстно-залежне уявлення $h_{k,j}^{\overleftarrow{LM}}$ токена t_k за даними токена $(t_{k+1}, \dots,\ t_N)$.

Двонаправлена мовна модель поєднує в собі як пряму, так і зворотну мовну моделі. Це формулювання максимізує логарифмічну ймовірність прямого і зворотного напрямків та обчислюється за формулою 2.4.

$$\sum_{k=1}^{N} (logp\left(t_k \middle| t_{1,} \dots t_{k-1}; \theta_x, \underrightarrow{\theta_{LSTM}},\ \theta_s\right) + \log p(t_k|t_{k+1}, \dots,\ t_N; \theta_x, \underleftarrow{\theta_{LSTM}}, \theta_s)) \tag{2.4}$$

Для прямого і зворотного мовних моделей параметри (θ_x) і шар Softmax (θ_s) об’єднуються, але в той же час всі інші параметри знаходяться окремо.

2.5.2 Тонке налаштування двонаправленої мовної моделі

В наступному експерименті будемо використовувати дискримінаційне тонке налаштування і поступове розморожування, техніку для збереження попередніх знань і уникнення катастрофічного забування при тонкій настройці.

Оскільки різні шари кодують різні типи інформації [11], їх слід налаштовувати по-різному. Для цього ми використовуємо дискримінаційну тонку настройку. Замість того, щоб використовувати ту ж швидкість навчання для всіх шарів моделі, дискримінаційна тонка настройка дозволяє налаштовувати кожен рівень з різними швидкостями навчання. Регулярне оновлення стохастичним градієнтним спуском (SGD) параметрів моделі θ на часовому кроці t обчислюється за формулою 2.5.

$$\theta_t = \theta_{t-1} - \eta \cdot \triangle_\theta j(\theta), \tag{2.5}$$

де η - швидкість навчання, $\triangle_\theta j(\theta)$ – градієнт цільової функції

Для дискримінаційної тонкої настройки потрібно розбити параметри θ на $\{\theta^1, \dots, \theta^L\}$, де θ^L містить параметри моделі на l- шарі, а L – кількість шарів моделі. Отже, аналогічно можемо отримати, що $\{\eta^1, \dots, \eta^L\}$, де η^1 – швидкість навчання l-шару.

SGD оновлення з дискримінаційною настройкою можна описати формулою 2.6.

$$\theta_t^l = \theta_{t-1}^l - \eta^l \cdot \triangle_{\theta l} j(\theta), \tag{2.6}$$

де η - швидкість навчання, $\triangle_{\theta l} j(\theta)$ – градієнт цільової функції з налаштуванням

Замість того, щоб тонко налаштовувати всі шари одночасно, що ризиковано через катастрофічне забуття, ми використовуємо поступове розморожування моделі, починаючи з останнього шару, оскільки він містить найменше спільне

знання [12]. Спочатку ми розморожуємо останній шар і тонко налаштовуємо все незамерзаючих шари за одну епоху. потім ми розморожуємо наступний нижній заморожений шар і повторюємо до тих пір, поки не закінчимо всі шари. Це схоже на «chain-thaw» [13], за винятком того, що ми додаємо шар кожного разу до безлічі «відтавати» шарів, а не навчаємо кожен шар окремо.

2.5.3 Глибинне контексто-залежне уявлення слів

Глибинні контекстно-залежні уявлення слів виходять з комбінації виходів проміжних шарів двонаправленої мовної моделі (ELMo embeddings). Поєднання внутрішніх станів дає дуже багаті репрезентації слів. Останні шари двонаправленої мовної моделі кодують контекстно-залежні смислові значення слова, в той час як більш ранні шари моделюють синтаксис (наприклад, вони можуть бути використовується для розмітки за частинами мови).

Одночасне об'єднання всіх цих сигналів дуже корисно. Для кожної конкретної задачі дозволяє вибирати найбільш кращу комбінацію.

Для кожного токена t_k, двонаправлена мовна модель із L шарами обчислює множину , що складається із $2L + 1$ векторних уявлень. Це можна відобразити формулою 2.7.

$$R_k = \left\{ x_k^{LM}, \overrightarrow{h_{k,j}^{LM}}, \overleftarrow{h_{k,j}^{LM}} \mid j = 1, \dots, L \right\} = \left\{ h_{k,j}^{LM} \mid j = 0, \dots, L \right\}, \tag{2.7}$$

де $h_{k,0}^{LM}$ – шар уявлення токена та $h_{k,j}^{LM} = [\overrightarrow{h_{k,j}^{LM}}; \overleftarrow{h_{k,j}^{LM}}]$, для кожного bi-LSTM шару.

Для включення в наступну модель, ELMo трансформує R в один вектор, $ELMok = E\ (Rk; \Theta e)$. У найпростішому випадку, ELMo просто вибирає верхній шар, $E\ (Rk) = h_{k,L}^{LM}$, як в TagLM [14] і CoVe [15]. У більш загальному сенсі, векторне подання слів виходить з усіх шарів двонаправленої мовної моделі та обчислюється за формулою 2.8.

$$ELMo_k = E(R_k; \theta) = \gamma \sum_{j=0}^{L} s_j h_{k,L}^{LM}, \quad (2.8)$$

де γ- скалярний параметр оптимізації, $h_{k,L}^{LM}$ – верхній шар

2.6 Архітектура двонаправленої мовної моделі

У цій роботі ми будемо використовувати AllenNLP tensorflow-реалізацію двонаправленої мовної моделі. Ця модифікована для підтримки спільного навчання обох напрямків і додавання залишкового з'єднання (residual connection) між шарами LSTM. Архітектуру мережі можна побачити на рисунку 2.4.

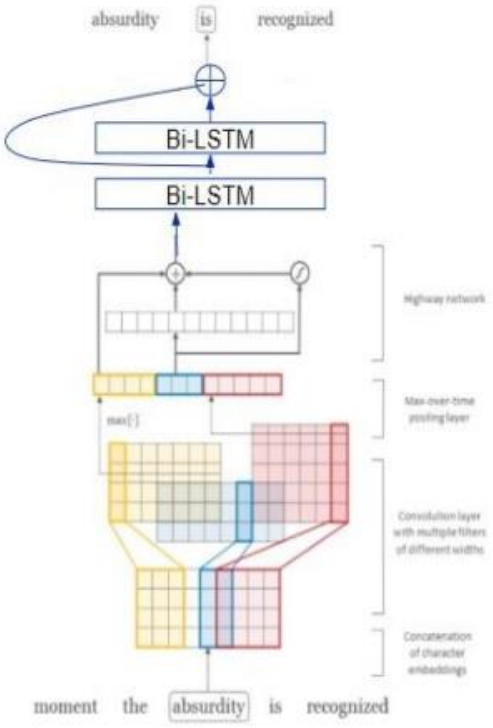

Рисунок 2.4 – Архітектура нейронної мережі, яка використовується у двонаправленій мовній моделі

Щоб збалансувати якість моделі з розміром моделі і обчислювальними вимогами для наступних завдань при збереженні чисто вхідного уявлення на основі символів, були змінені всі вхідні і внутрішні розмірності в кращій моделі CNN-BIG-LSTM. Кінцева модель використовує $L = 2$ шари bi-LSTM розмірністю 2048 і 512 розмірними проекціями. Також використовується residual connection між першим і другим шаром.

Контекстно-незалежне уявлення використовує 2048 згорткових фільтрів, за якими слід два highway шару і лінійна проекція до 512 розмірного уявлення.

2.7 Архітектура рекурентної нейронної мережі з багатошаровим перцептором

Для початкового рішення задачі розпізнавання іменованих сутностей була спроектована архітектура рекурентної нейронної мережі з багатошаровим перцептором. Було взято 25 різних морфологічних і виділених вручну ознак (лема, частина мови, префікс, постфікси, грамема, на якій позиції стоїть в реченні, чи є першим словом в пропозиції, чи є останнім і інші). Відповідно до побудованим на кожному реченні синтаксичним деревом ми доповнили уявлення слова, поданням предка і предка предка. Також доповнили ознаками оточуючих слів з вікном в п'ять слів. Архітектуру нейронної мережі можна побачити на рисунку 2.5.

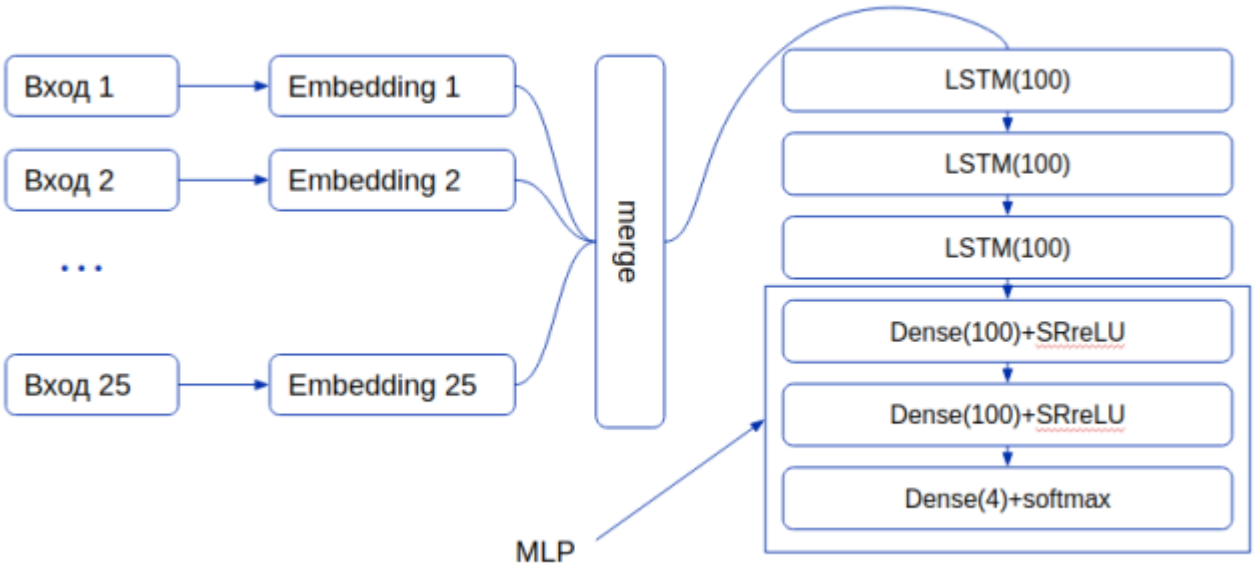

Рисунок 2.5 – Рекурентна нейронна мережа + МСП

У підсумку вийшло 825 розмірне уявлення, яке йде на вхід рекурентної нейронної мережі, що складається з трьох шарів LSTM і трьох повнозв'язних шарів (багатошаровий перцептор). Також застосовується варіаційний дропаут для запобігання перенавчання та досягнення кращих результатів.

2.8 Архітектура двонаправленої рекурентної нейронної мережі

Слідуючи останнім новітнім системам, базова модель для рішення задачі розпізнавання іменованих сутностей використовує попередньо навчені уявлення слів, два шари bi-LSTM, повнозв'язний шар, виходи з якого йдуть на вхід алгоритму умовних випадкових полів (CRF). Два шару bi-LSTM, перший з яких містить 40 нейронів, а другий - 20 нейронів, повнозв'язну шар містить число нейронів дорівнює кількості тегів (число іменованих сутностей + 1). Також застосовується варіаційний дропаутов для запобігання перенавчання та досягнення кращих результатів. Архітектуру нейронної мережі можна побачити на рисунку 2.6

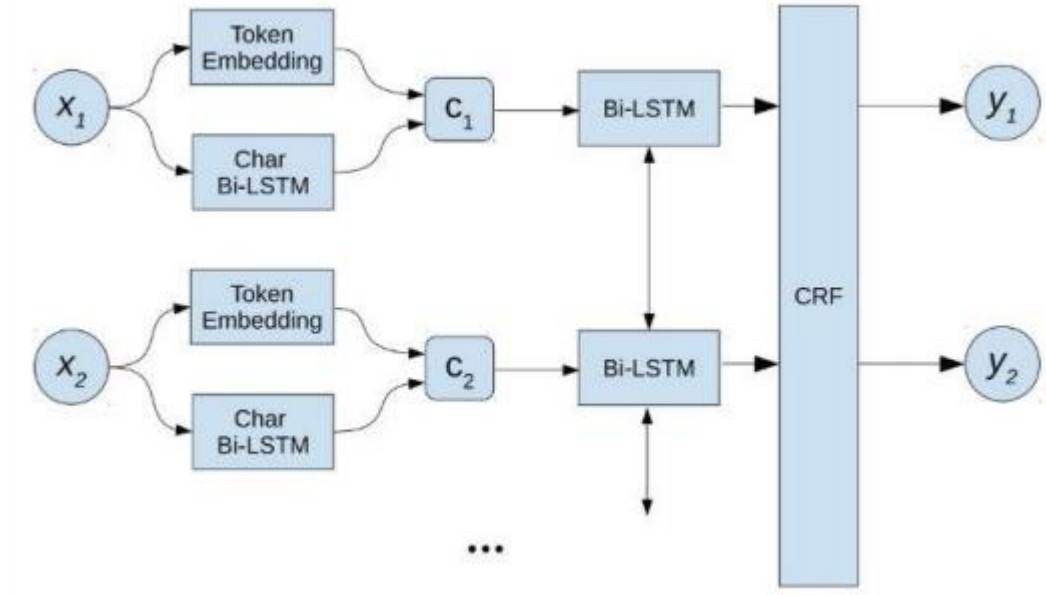

Рисунок 2.6 – Архітектура двонаправленої рекурентної нейронної мережі із CRF шаром

Під час тестування описаної архітектури виконується декодування з використанням алгоритму Вітербо.

3 ОПИС ЕКСПЕРИМЕНТАЛЬНИХ ДОСЛІДЖЕНЬ

Всі практичні дослідження проводилися на мові Python 3.8. для проведення експериментів і реалізації спроектованих архітектур нейронних мереж використовувалися бібліотеки машинного навчання Keras і Tensorflow. Час навчання нейронних мереж залежить від розміру даних і розміру самої нейронної мережі. В роботі, наприклад, двонаправлена рекурентна нейронна мережа з CRF шаром навчалася на 30 епохах з використанням одного GPU протягом години, а ось двонаправлена мовна модель з використанням трьох GPU навчається близько 10 днів.

3.1 Набір даних

У роботі використаний FactRuEval корпус даних. Корпус складається з новинних і аналітичних текстів російською мовою, присвячених соціальних і політичних питань. Тексти були зібрані з таких джерел:

– private Correspondent (http://www.chaskor.ru/);

– wikinews (https://ru.wikinews.org).

Корпус був розділений на дві частини: демонстраційний корпус з 122 текстів і тестовий корпус з 133 текстів. Демонстраційний корпус має близько 30 тисяч токенов і 1700 пропозицій. Тестовий корпус має близько 60 тисяч токенов і 3100 пропозицій. Моделі нейронних мереж для вирішення завдання розпізнавання іменованих сутностей навчалися на демонстраційному корпусі і тестувалися на тестовому корпусі. Необхідно було класифікувати кожен токен на наступні класи: person, location, organization, інші (O).

3.2 Методи оцінки якості розпізнавання іменованих сутностей

Оцінка якості роботи системи розпізнавання іменованих сутностей проводиться на текстових корпусах, розміщених експертами вручну; при цьому способи вимірювання якості можуть змінюватися.

NER по суті включає дві підзадачі: виявлення кордонів та ідентифікація типу. У "оцінці точної відповідності", правильно розпізнаний екземпляр вимагає систему правильно визначити його кордони та тип одночасно.

Більш конкретно, кількість хибно-позитивні (False positives, FP), хибно-негативні (False negatives, FN) та істинно-позитивні (True positives, TP) використовуються для того, щоб підрахувати точність (Precision), повноту (Recall) та F-міру:

- хибно -позитивні (False positives, FP) : сутність, яку повертає NER, але не насправді вона відсутня в контексті;
- хибно -негативні (False negatives, FN) : сутність, яку не повертає система NER, але насправді вона присутня в контексті;
- істинно - позитивні (True positives, TP) : сутність, яку повертає система NER і вона присутня в контексті.

Точність стосується відсотка результатів розробленої системи, які правильно розпізнаються. Повнота стосується відсотка від загальної кількості визнаних розробленою системою об'єктів. Точність та повнота обчислюються за формулами 3.1 та 3.2

$$Precision = \frac{TP}{TP+FP}, \tag{3.1}$$

$$Recall = \frac{TP}{TP+FN}, \tag{3.2}$$

де TP – істинно-позитивне рішення, FP – хибно-позитивне рішення, FN – хибно-негативне рішення

Мірою, що поєднує точність і відкликання, є гармонійне середнє значення точності та відкликання, традиційний F-міра або збалансований F-бал та обчислюється за формулою 3.3.

$$F = 2\frac{Precision * Recall}{Precision + Recall} \quad (3.3)$$

Крім того, макроусереднене F-бальне значення та мікроусереднене F-бальне значення враховують ефективність для кількох типів сутності. Макроусереднений F-бал незалежно обчислює F-бал за різними типами сутності, а потім бере середнє значення F-балів. Сума мікроусереднених F-балів до окремих помилкових негативів, помилкових позитивних і істинних позитиви для всіх типів сутності потім застосовують їх для отримання статистики. Останні можуть сильно вплинути на якість визнання сутностей у великих класах у корпусі.

Також на конференції було запропоновано розмічений набір даних, що складається з колекції документів, призначеної для навчання, і колекції документів, призначеної для тестування. Дані CoNLL'03 часто використовуються дослідниками для оцінки якості систем розпізнавання іменованих сутностей.

Існують і інші способи оцінки якості. Так, на конференціях MUC якість роботи систем вимірювалося за двома «осями»: здатності правильно розпізнати тип і здатності виділити правильні межі сутності. Цей спосіб оцінки дозволяє більш поблажливо ставитися до таких помилок, як невірно визначені кордони сутності (якщо її тип був визначений правильно) і невірно визначений тип сутності (якщо її кордони були визначені правильно). На конференції ACE був запропонований складний спосіб оцінки якості, що дозволяє враховувати в оцінці значимість одних типів помилок над іншими. Однак це проблематично, оскільки підсумкові бали можна порівняти лише за параметрами, які закріплені. Складні методи оцінювання не є інтуїтивно зрозумілими і ускладнюють аналіз помилок.

Таким чином, складні методи оцінки не широко використовуються в останніх дослідженнях.

3.3 Навчання нейронної мережі на морфологічних та синтаксичних ознаках

Початковим припущенням було те, що для вирішення завдання розпізнавання іменованих сутностей на вхід рекурентній нейронній мережі необхідно подати таке подання слів, яке містить в собі морфологічні та синтаксичні ознаки. Векторне подання слова будувалося на основі ознак слів, що стоять поруч, які є предками в синтаксичному дереві розбору. У підсумку виходить 33 слова:

- '' - поточний слово
- '2dom' - батько батька поточного слова
- '1dom' - батько поточного слова
- '-5' - 5е слово зліва від поточного («-» - зліва, «+» - праворуч)
- '-5_2dom' - батько батька 5-го слова зліва від поточного
- '-5_1dom' - батько «-5го»

Також вийшло 25 різних ознак для кожного слова: prefix4 prefix3 prefix2 prefix1 postfix4 postfix3 postfix2 postfix1 posStart pos link len lemma isupper istitle islower isdigit isalpha isalnum isLastWord isFirstWord grm forma dom ID.

Як зазначалося в пункті 2.7 всього на вхід виходить 825 ознак. Архітектура мережі описана також у пункті 2.7. Повний набір параметрів для цієї моделі складається з параметрів кожного шару нейронної мережі (вагові матриці, зміщення, матриці векторних уявлень слів). Всі ці параметри налаштовуються під час тренування за алгоритмом зворотного поширення помилки зі стохастичним градієнтним спуском (lr = 0.1, decay = 1e-7, momentum = 0, clipvalue = 3). Після навчання 30 епох модель RNN + MLP (пункт 2.7) досягає precision - 72.90%, recall - 76.60% і F1 - 74.71%. Вхідні ознаки кодують недостатньо добре знання природної мови, і модель навчається на невеликій кількості розмічених даних. Для кращих результатів необхідно глибоке розуміння граматики, семантики і інших елементів природної мови, чого немає в повній мірі у взятих ознаках.

3.4 Навчання двонаправленої рекурентної нейронної мережі з CRF шаром дистрибутивним векторним уявленням слів

Подальшим припущенням було додати на вхід уявлення слів, взяте з дистрибутивної семантичної моделі. Дистрибутивні семантичні моделі кодують смислове значення слів. Ці ознаки ми об'єднували з деякими морфологічними ознаками і подавали на вхід двонаправленій рекурентній нейронній мережі з CRF шаром (пункт 2.8). Як дистрибутивні векторні уявлення були взяті уявлення з дистрибутивної семантичної моделі fastText, що навчалися заздалегідь на текстах Common Crawl і Wikipedia. Ці моделі були навчені з використанням CBOW алгоритму, розмірністю 300, з літерними n-грамами довжини 5, вікном розміром 5. Повний набір параметрів для цієї моделі складається з параметрів шарів bi-LSTM (вагових матриць, зсувів, матриці векторних уявлень слів) і матриці переходу шару CRF.

Всі ці параметри налаштовуються під час тренування за алгоритмом зворотного поширення з стохастичним градієнтним спуском (lr = 0.1, decay = 1e-7, moemntum = 0, clipvalue = 3). Variational Dropout застосовується для запобігання перенавчання та поліпшення продуктивності моделі. Після навчання на 30 епохах модель bi-LSTM + CRF досягає precision - 77.23%, recall - 85.19% і F1 - 81.02%.

3.5 Навчання двонаправленої рекурентної нейронної мережі із CRF шаром та ELMo уявленням слів

Ми навчаємо модель biLM на великому корпусі даних. Архітектура мережі описана в пункті 2.6, в роботі використана AllenNLP tensorflow реалізація. Дані взяті з conll2017. Після підготовки даних виходить близько 200 мільйонів токенов і 15 мільйонів пропозицій, словник становить 60 тисяч найчастіше вживаних слів.

Після навчання протягом 5 епох перплексія становить близько 45. Після попереднього навчання biLM може обчислювати векторні уявлення слів для будь-якого завдання. Для подальшого навчання моделі двонаправленої рекурентної мережі для вирішення задачі розпізнавання іменованих сутностей були сформовані ElMo уявлення слів, отримані з biLM. Повна множина параметрів при навчанні мовної моделі:

```
options = {
'batch_size': 64,
'n_gpus': 3,
'bidirectional': True,
'char_cnn': {'activation': 'relu',
'embedding': {'dim': 16},
'filters': [[1, 32],
[2, 32],
[3, 64],
[4, 128],
[5, 256],
[6, 512],
[7, 1024]],
'max_characters_per_token': 50,
'n_characters': 261,
'n_highway': 2},
'dropout': 0.1,
 'lstm': {
'cell_clip': 3,
'dim': 4096,
'n_layers': 2,
'proj_clip': 3,
'projection_dim': 512,
'use_skip_connections': True},
'all_clip_norm_val': 10.0,
'n_epochs': 10,
'n_train_tokens': n_train_tokens,
'batch_size': batch_size,
'n_tokens_vocab': vocab.size,
'unroll_steps': 20,
'n_negative_samples_batch': 8192,
}
```

3.5.1 Тонке налаштування двонаправленої мовної моделі

У деяких випадках тонке налаштування biLM на даних, які стосуються задачі, призводить до значних падінь перплексії і збільшення продуктивності наступних моделей. Для тонкого налаштування моделі ми використовуємо нерозмічену пропозиції демонстраційного корпусу FactRuEval, які також були доступні під час змагання. BiLM навчається на 3 епохах і оцінюється за пропозиціями тестового корпусу. Перплексія на тестовому корпусі після тонкого налаштування впала з 90 до 40 (чим менше, тим краще).

Після тонкого налаштування biLM були сформовані ELMo уявлення слів, які подаються на вхід іншої моделі.

3.5.2 Результати із ElMo уявленням

Дистрибутивні векторні уявлення слів (fastText) об'єднуються з глибокими контекстно-залежними уявленнями (ELMo) до тонокого налаштування biLM і подаються на вхід bi-LSTM мережі з CRF шаром. Після навчання на 30 епохах модель Bi-LSTM + CRF досягає precision - 82.32%, recall - 84.0% 4i F1 - 83.17%.

В кінці дистрибутивні векторні уявлення слів (FastText) об'єднуються з глибокими контекстно-залежними уявленнями (ELMo) після тонокого налаштування biLM і подаються на вхід bi-LSTM мережі з CRF шаром. Після навчання на 30 епохах модель Bi-LSTM + CRF досягає precision - 83.19%, recall - 85.41% і F1 - 84.29%.

3.6 Порівняння результатів

Традиційні підходи до розпізнавання іменованих сутностей в російською мові в значній мірі грунтувалися на ручних правилах і зовнішніх ресурсах. Так застосовувалися регулярні вирази і словники для вирішення завдання. Наступним кроком було застосування методів статистичного навчання, таких як умовні випадкові поля (CRF) і векторні машини (SVM) для класифікації сутностей. CRF над лінгвістичними ознаками, розглянутими як базовий рівень в дослідженні.

Також пропонували двоетапний алгоритм умовних випадкових полів: на першому етапі на вхід для CRF подавалися вручну виділені лінгвістичні ознаки. Потім на другому ті ж лінгвістичні ознаки були об'єднані з глобальної статистикою, розрахованої на першому етапі і подавалися в CRF. В роботі [16] застосовували класифікацію SVM до дистрибутивним векторним уявленням слів і фраз. Ці уявлення були отримані шляхом навчання без учителя дистрибутивних семантичних моделей на різних новинних корпусах даних. Одночасне використання словникових ознак і розподілених векторних уявлень слів було представлено в [17]. словникові ознаки були вилучені з Wikidata, а векторні уявлення були попередньо предобучені на даних з Вікіпедії. Потім ці об'єднані ознаки використовувалися для класифікації за допомогою SVM. В цей момент методи навчання з використанням глибоких нейронних мереж є найбільш перспективним варіантом для NER. В роботі [18] запропонована глибока нейронна мережа LSTM для вирішення задачі розпізнавання іменованих сутностей російською мовою. Відмінною особливістю цієї роботи є зв'язування задачі моделювання мови з класифікацією іменованих сутностей, що дуже схоже на підхід, який використовується в даної випускний роботі. У дослідженні [19] застосовується сучасна модель нейронної мережі для англійської NER до відкритих розмічених для NER російською мовою корпусам даними. Модель складається з трьох основних компонентів: двонаправлені рекурентні мережі (bi-LSTM), CRF і розподілене векторне подання слів. Експерименти показують, що тільки Bi-LSTM

дещо гірше, ніж модель с CRF [20]. Додавання CRF в якості наступного етапу обробки поверх шарів Bi-LSTM значно покращує продуктивність моделі і дозволяє перевершити модель, представлену в [20]. Різниця моделі bi-LSTM + CRF моделі [19], від моделі представленої в [20], є навчені векторні уявлення слів і фраз. навчання мережі Bi-LSTM ще й на рівнях символів дає кращі результати, ніж ручні ознаки в [20]. Як уже зазначалося дистрибутивні векторні уявлення стали стандартним інструментом в області обробки природної мови. Такі уявлення здатні захоплювати семантичні особливості слів і значно покращувати результати для різних завдань, що продемонстровано в роботі [19]. Автори роботи [19] вважають, що результати LSTM highway network можуть бути поліпшені шляхом кращої ініціалізації ваг і глибшої архітектурою. Крім того, укладання шарів highway LSTM може поліпшити результати, що дозволяють мережі динамічно коригувати складність обробки. Результати всіх описаних підходів на корпусі даних FactRuEval показані в таблиці 1.

Таблиця 1 – Порівняння результатів дослідження

	Precision	Recall	F1-score
Wiki-based-aproach [17]	88.19	64.75	74.67
basic+dictionary + w2v features on SVM [17]	82.57	74.08	78.10
SVM+w2v [16]	-	-	82.88
Bi-LSTM + CRF + Lenta [19]	83.80	80.84	82.10
NeuroNER + Highway char [19]	80.59	80.72	80.66

Кінець таблиці 1

	Precision	Recall	F1-score
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

У цьому дослідженні було виявленно, що контекстно-залежне уявлення, корисно в задачі розпізнавання іменованих сутностей в текстах на російській мові. Коли було налаштовано двонаправлену мовну модель і включено такі уявлення, ми досягли найкращих результатів. Також в дослідженні показано підвищення точності моделі з поступовою її адаптацією, додаванням різних знань морфології, синтаксису і семантики.

Надалі планується більш ретельна робота над навчанням мовної моделі.

ВИСНОВКИ

Під час проведення дослідження у межах атестаційної роботи було проведено аналіз використання методів розпізнавання іменованих сутностей у новинному контенті на російській мові.

У даній роботі була досягнуті найкращі результати у вирішенні завдання розпізнавання іменованих сутностей в російській мові. Саме завдання в світі вже вирішується давно і на різних мовах. В англійській мові зараз найвищі результати. За два десятки років тому починали вирішувати завдання з допомогою рукописних правил, поступово ускладнюючи алгоритми, застосовуючи статистичні моделі. Зараз у всіх мовах з цим завданням краще за все справляються глибокі двонаправлені рекурентні нейронні мережі. Їх успіх продиктований багатьма факторами. Щодо завдання розпізнавання іменованих сутностей можна виділити те, що для об'єкта дійсно важливий контекст. Для нейронної мережі все як і раніше важливо для навчання мати великий набір даних.

У даній роботі були проаналізовані існуючі підходи до розв’язання задачі именінованних сутностей, а також представлений метод навчання з частковим залученням вчителя, коли інша модель навчається на великому корпусі нерозмічену даних і знання від неї можна передати моделі, яка буде вирішувати конкретну задачу, але навчатися на невеликій кількості розмічених даних. В кінці було проведено експериментальне завдання щодо порівняння отриманих результатів розробленого підходу з вже існуючими.

Це була двунаправленная мовна модель, яка вивчає граматику мови. Раніше було показано, що на різних їй шарах кодуються різні знання. На початку роботи наводилася піраміда завдань в природній мові. Часто в рішенні доменних завдань бувають проблеми з присутністю специфічних слів і виразів, яких не було при навчанні моделей без вчителя. Для боротьби з такими словами мовну модель навчають, наприклад, на частинах слова або на буквах. Також застосовують тонке налаштування нейронної мережі, що витягти знання з доменної області. Однак,

тонке налаштування необхідно проводити акуратно, щоб надалі доменна модель не перенавчалася і зберігала узагальнюючу здатність.

У російській мові є ще напрямки, в яких можна розвивати рішення задачі розпізнавання іменованих сутностей. Можно продовжити дослідження в напрямку мовних моделей.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАНЬ

1. Safavian S. R., Landgrebe D. A survey of decision tree classifier methodology//IEEE transactions on systems, man, and cybernetics. – 1991. – Т. 21. – №. 3. – С. 660-674.
2. Krogh A. et al. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes1 //Journal of molecular biology. – 2001. – Т. 305. – №. 3. – С. 567-580.
3. Chapelle O., Scholkopf B., Zien A. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews] //IEEE Transactions on Neural Networks. – 2009. – Т. 20. – №. 3. – С. 542-542.
4. Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, p. 436, 2015.
5. D. Nadeau and S. Sekine, “A survey of named entity recognition and classification,” Lingvist. Investig., vol. 30, no. 1, pp. 3–26, 2007.
6. . Bengio Y., Simard P., Frasconi P. Learning long-term dependencies with gradient descent is difficult //IEEE transactions on neural networks. – 1994. – Т. 5. – №. 2.– С. 157-166.
7. Hochreiter S., Schmidhuber J. Long short-term memory //Neural computation. – 1997. – Т. 9. – №. 8. – С. 1735-1780.
8. Schuster M., Paliwal K. K. Bidirectional recurrent neural networks //IEEE Transactions on Signal Processing. – 1997. – Т. 45. – №. 11. – С. 2673-2681.
9. Lafferty J., McCallum A., Pereira F. C. N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. – 2001.
10. Pennington J., Socher R., Manning C. Glove: Global vectors for word representation //Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). – 2014. – С. 1532-1543.
11. Howard J., Ruder S. Fine-tuned Language Models for Text Classification //arXiv preprint arXiv:1801.06146. – 2018.

12. Yosinski J. et al. How transferable are features in deep neural networks? //Advances in neural information processing systems. – 2014. – С. 3320-3328.
13. Felbo B. et al. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm //arXiv preprint arXiv:1708.00524. – 2017
14. Peters M. E. et al. Semi-supervised sequence tagging with bidirectional language models //arXiv preprint arXiv:1705.00108. – 2017.
15. McCann B. et al. Learned in translation: Contextualized word vectors //Advance in Neural Information Processing Systems. – 2017. – С. 6297-6308.
16. Ivanitskiy R., Shipilo A., Kovriguina L. Russian Named Entities Recognition and Classification Using Distributed Word and Phrase Representations //SIMBig. – 2016. – С. 150-156.
17. Сысоев А. А., Андрианов И. А. Распознавание именованных сущностей: подход на основе вики-РесуРсов. 2016
18. Malykh V., Ozerin A. Reproducing Russian NER Baseline Quality without Additional Data //CDUD@ CLA. – 2016. – С. 54-59.
19. Anh L. T., Arkhipov M. Y., Burtsev M. S. Application of a Hybrid Bi-LSTM-CRF model to the task of Russian Named Entity Recognition //arXiv preprintarXiv:1709.09686. – 2017.
20. Gareev R. et al. Introducing baselines for Russian named entity recognition //International Conference on Intelligent Text Processing and Computational Linguistics. – Springer, Berlin, Heidelberg, 2013. – С. 329-342.
21. T. Shatovska, I. Kamenieva and I. Tarasov, "New module of text classification for IDA system," 2009 10th International Conference - The Experience of Designing and Application of CAD Systems in Microelectronics, Lviv-Polyana, 2009, pp. 481-482.
22. Валенда Н.А., Самофалов Л.Д. Функционально-семантический анализ конструкций естественного языка на основе унификации // Системи обробки інформації. – 2016. – №7 (144). – С. 79-82.