NLP: опрацювання омонімів в українськомовних текстах

Доповідач: Парнак Д. А.
Керівник: ст. в. Смиш О. Р.

2024

Обґрунтування вибору теми дослідження

Брак готових рішень

Неоднозначність омонімів

Неправильне опрацювання на морфологічному рівні

Помилки в результатах роботи систем із використанням NLP

Мета роботи

уможливити підвищення точності розомонімізації в українській мові, шляхом створення системи для розпізнавання та розбору омонімів в текстах, написаних українською мовою.

Об'єкт дослідження

Оброблення текстових даних, написаних природною українською мовою.

Предмет дослідження

Розроблення системи для опрацювання омонімів, а також тренування моделі розпізнавання омонімів з метою використання в проєктах з обробки текстових даних.

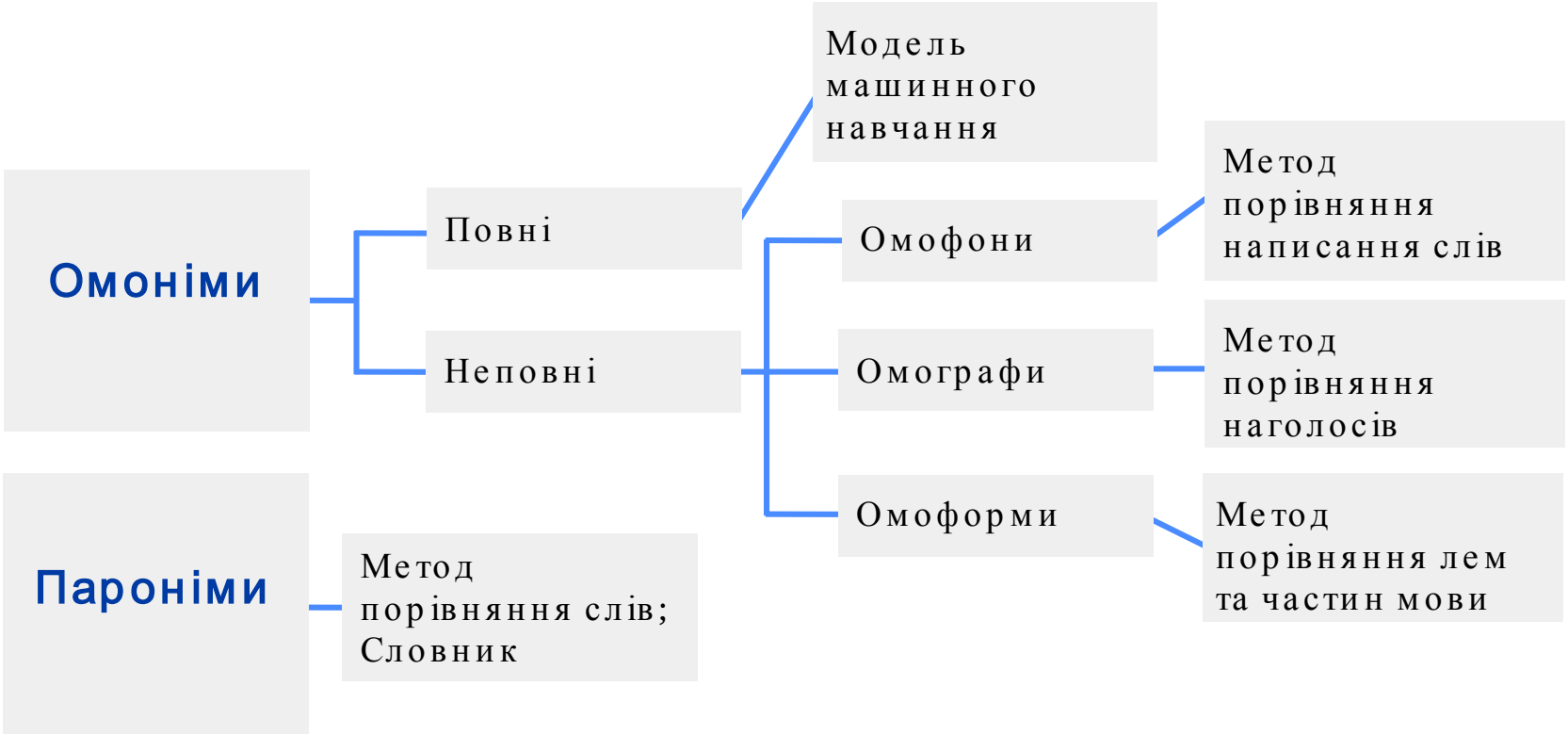
Завдання роботи

- аналіз поточних рішень для обробки природної української мови, зокрема розпізнавання омонімів в текстах;
- формування методів та тренування моделі машинного навчання для розпізнавання та аналізу різних груп омонімів та паронімів;
- апробація результатів застосування реалізованих методів.

Наукова новизна

- уперше проаналізовано наявні рішення для автоматизованої роботи з українськомовними омонімами в тексті;
- уперше розроблено комплексну систему для автоматизованого виявлення та розбору всіх груп омонімів із використанням методів на основі правил та моделі машинного навчання.

Аналіз омонімів та паронімів



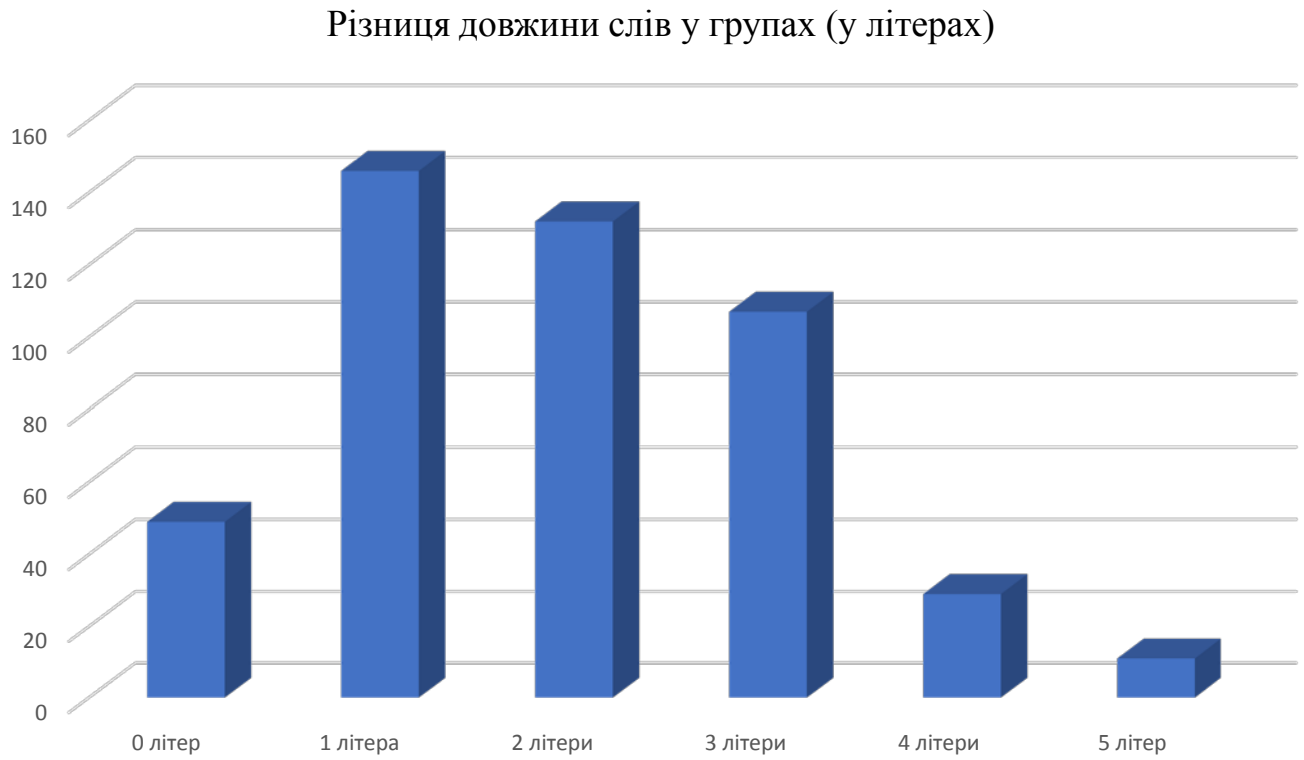
Аналіз словника паронімів

paronyms_json

```
{
   "group": "КОНТАКТ // КОНТРАКТ",
   "paronym1": "контакт",
   "def1": "Діловий зв'язок, тісні стосунки, взаємна узгодженість дій.",
   "paronym2": "контракт",
   "def2": "Письмова угода, договір із взаємними зобов'язаннями його учасників."
}
```

Аналіз словника паронімів

Аналіз словника паронімів



Розроблені методи

Опрацювання омоформ

- Очищення тексту від пунктуації
- Встановлення лем, та частин мови лексем та їхнє паралельне порівнювання

Опрацювання омографів

- Встановлення подвійних наголосів
- Підтвердження істинності омографів за врахування розбіжності тлумачень

Опрацювання омофонів

- Очищення тексту від пунктуації
- Пошук омофонів за регулярними виразами

Опрацювання паронімів

- Пошук паронімічних груп за регулярними виразами
- Виокремлення окремих паронімів та тлумачень зі словника

Приклад роботи вебінтерфейсу програми

Результати аналізу

Оригінальний текст:

Скоро розпочинається вступна кампанія. Ці брати були дружніми. Цей замок обабіч місця, де ми часто збиралися разом; там завжди яскраво світило сонце. Не можна брати цей атлас, він не мій. Якби ж я міг мати щось більше, ніж просто те, що дала мені дорога мати. Сон - це найкращий відпочинок. Я би не хотів працювати у цій компанії.

Аналіз:

Омоформи:
брати - леми: ['брат', 'брати'], частини мови: ['NOUN', 'VERB']
мати - леми: ['мати', 'мати'], частини мови: ['VERB', 'NOUN']

Слова з різним наголосом:
замок
1: ЗАМОК, мка, ч. Укріплене житло феодала доби середньовіччя з оборонними, господарськими, культовими і т. ін. будівлями, звичайно оточене високим кам'яним муром з кількома вежами.
2: ЗАМОК, мка, ч. Пристрій для замикання дверей у приміщеннях, дверцят шафи, скринь, шухляд і т. ін.

атлас
1: АТЛАС, у, ч. Укладений за певною системою і виданий у формі альбома або книжки збірник географічних, історичних та ін. карт.
2: АТЛАС, у, ч. Шовкова або напівшовкова тканина, блискуча і гладенька з лиця.

Омофони:
Сон - це - сонце.

Пароніми (зі словника):
кампанія паронім до компанія

Пароніми (в тексті):
кампанія - компанії

Тренування моделі

- Омонім: ”ручка”
- 504 речення в датасеті

Формування датасету: GPT4

Препроцесинг UDPipe

Векторизація TF-IDF

Тренування моделі логістиної регресії

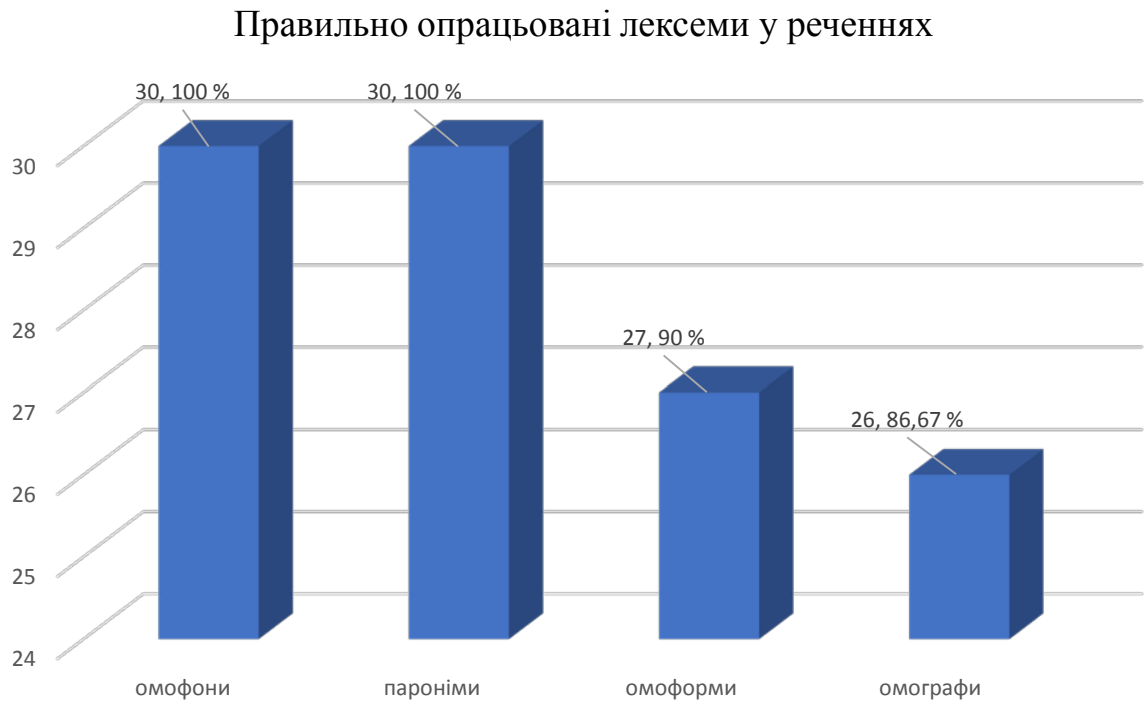
Результати тренування моделі

classification_report

	precision	recall	f1-score	support
Зменшувальна до рука	0.87	0.93	0.90	50
Прилад для писання	0.93	0.93	0.93	50
Частина предмета	0.92	0.86	0.89	50
accuracy			0.90	150
macro avg	0.91	0.90	0.90	150
weighted avg	0.91	0.90	0.90	150

Accuracy: 0.9047619047619048

Апробація результатів: аналіз точності методів на основі правил



Апробація результатів: аналіз точності моделі машинного навчання

Омонім: ”вид”

classification_report

	precision	recall	f1-score	support
Вигляд, краєвид	0.92	0.88	0.90	13
Окрема галузь, категорія	0.93	0.90	0.91	13
accuracy			0.90	26
macro avg	0.92	0.89	0.90	26
weighted avg	0.92	0.89	0.90	26

Accuracy: 0.90354630104

Омонім: ”нота”

classification_report

	precision	recall	f1-score	support
Музичний знак	0.93	0.91	0.92	13
Офіційне звернення	0.92	0.91	0.91	13
accuracy			0.92	26
macro avg	0.93	0.91	0.92	26
weighted avg	0.93	0.91	0.92	26

Accuracy: 0.923890416428

Омонім: ”дисципліна”

classification_report

	precision	recall	f1-score	support
Підтримання порядку	0.91	0.90	0.90	13
Предмет, розділ науки	0.88	0.89	0.88	13
accuracy			0.89	26
macro avg	0.89	0.90	0.89	26
weighted avg	0.89	0.90	0.89	26

Accuracy: 0.8924538022519

Висновки

- Здійснено аналіз ресурсів, які застосовуються для автоматичного опрацювання омонімів в обробці природної української мови;
- Проведено аналіз проблематики омонімії та паронімії в контексті україномовного NLP;
- Розроблено методи на основі правил для розпізнавання часткових омонімів;
- Натреновано модель розпізнавання повного омоніма з точністю 90,47 %, що може використовуватись як попередньо тренована
- Здійснено апробацію результатів роботи, встановлено точність розробленої системи на рівні понад 86 %

Дякую за увагу!