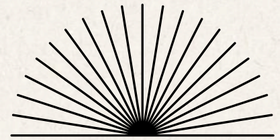
PROBLEMS IN UKRAINIAN NLP

Катерина Мудра, КН4



Мета

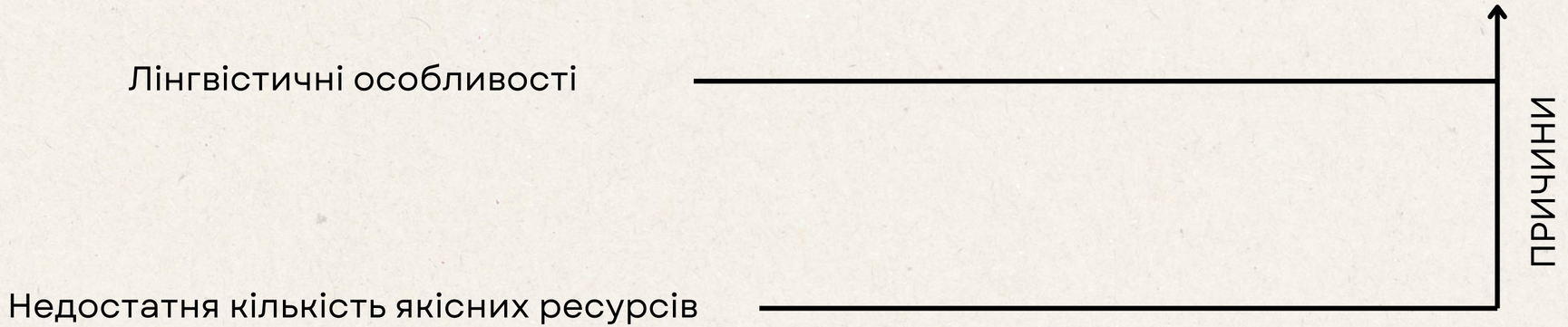
дослідження основ обробки природної мови (NLP) з фокусом на застосування цієї технології для обробки української мови.

а саме:

- аналіз наявних лінгвістичних ресурсів та проблем, що виникають при обробці українського тексту
- віднайдення шляхів вирішення цих проблем для подальшого вдосконалення NLP-систем

Актуальність роботи

значні виклики у створенні точних і ефективних моделей для обробки українського тексту



Структура

Огляд NLP #1

основи NLP, зокрема концепції, що охоплюють Natural Language Understanding (NLU) і Natural Language Generation (NLG), а також основні напрямки і застосування цих технологій у різних сферах

Стан NLP в Україні #2

Історія розвитку NLP в Україні, починаючи з наукових досягнень в 1960-х роках і до сучасних успіхів у створенні лінгвістичних ресурсів та моделей для української мови.

Проблеми в українському NLP #3

специфічні проблеми, з якими стикається обробка української мови; шляхи подолання цих проблем та можливості для подальшого розвитку NLP в Україні.

NLU?

NLP - це

підгалузь штучного інтелекту, яка являє собою комп’ютеризований підхід до розпізнання, обробки та аналізу людської мови.

NLG?

Задачі NLP

Низькорівневі

Токенізація
Нормалізація
Лематизація/
стемінг
POS-розмітка

Середньорівневі

Парсинг
Chunking
Аналіз
залежностей

Високорівнені

Машинний переклад
Sentiment analysis
Підсумовування тексту
Генерування тексту
Діалогові системи

Складові аналізу

Складова аналізу	Опис	Інструменти
Фонетичний аналіз	Аналіз звуків мови, їх характеристик та відповідність написанню	- eSpeak (має підтримку української) - Festival TTS (обмежена укр. підтримка)
Фонологічний аналіз	Дослідження фонем, наголосу, інтонації тощо	- Частково підтримується в TTS/ASR системах - Дослідницькі моделі (академічні)
Морфологічний аналіз	Визначення частин мови, граматичних форм (рід, число, відмінок тощо)	- БРУК - ГРАК - pymorphy2 (для рос., з адаптацією) - LangTool API
Лексичний аналіз	Розпізнавання слів, токенізація, нормалізація	- Stanza (від Stanford NLP, має укр. модель) - SpaCy (з кастомним pipeline)

Складові аналізу

Складова аналізу	Опис	Інструменти
Синтаксичний аналіз	Визначення граматичних зв’язків між словами (дерева залежностей)	- UDPipe - Stanza - SyntaxNet (обмежено для укр.)
Семантичний аналіз	Визначення значення слів, синонімії, контексту	- Word2Vec для укр. (Languk) - FastText від Facebook - BERT-Ukraine (UBERT)
Прагматичний аналіз	Розуміння наміру мовця, контексту ситуації	- ChatGPT/OpenAI API (може працювати з укр. запитами) - Rule-based системи
Дискурсний аналіз	Аналіз зв’язності тексту, логіки переходів, тем	- Частково реалізовано в LangTool NLP - GPT-моделі для укр. тексту

Мовні особливості

1	Флективність та багата морфологія
2	Вільний порядок слів
3	Фонетичні особливості та наголос
4	Синонімія, варіативність та стилістичне багатство
5	Складна система префіксів і суфіксів
6	Кальки та запозичення
7	Діалекти та регіональні варіанти

Проблеми корпусів

Розмітка тексту
Репрезентативність корпусу
Подання результатів
Веб як джерело корпусу

Проблеми NLP для розробників

Машини погано справляються з багатозначністю, синонімією, омонімією, метафорами, грою слів і контекстом
Проблема визначення емоцій і тональності, складність у виявленні емоційного тону повідомлення
Кросмовна морфологія і малоресурсні мови
Проблема гумору і креативності

Шляхи подолання

Наповнення поточних корпусів більшою кількістю даних
Візуалізація та інтерпретація даних
Подолання багатозначності й контекстної неоднозначності.
Розвиток емоційного та прагматичного аналізу
Підтримка малоресурсних мов та діалектів
Розпізнавання гумору та креативності