Система генерації текстів на основі методів нейронних мереж

студент 6 курсу
КА-61м, Лаврій Богдан

Національний технічний університет України
"Київський політехнічний інститут імені Ігоря Сікорського"
Інститут прикладного системного аналізу

керівник: доцент кафедри ММСА, к.т.н., доцент
Тимощук О. Л.

Об’єкт дослідження

Методи обробки природної мови

Предмет дослідження

Штучні нейронні мережі

Актуальність

Генерація текстів потрібна для:

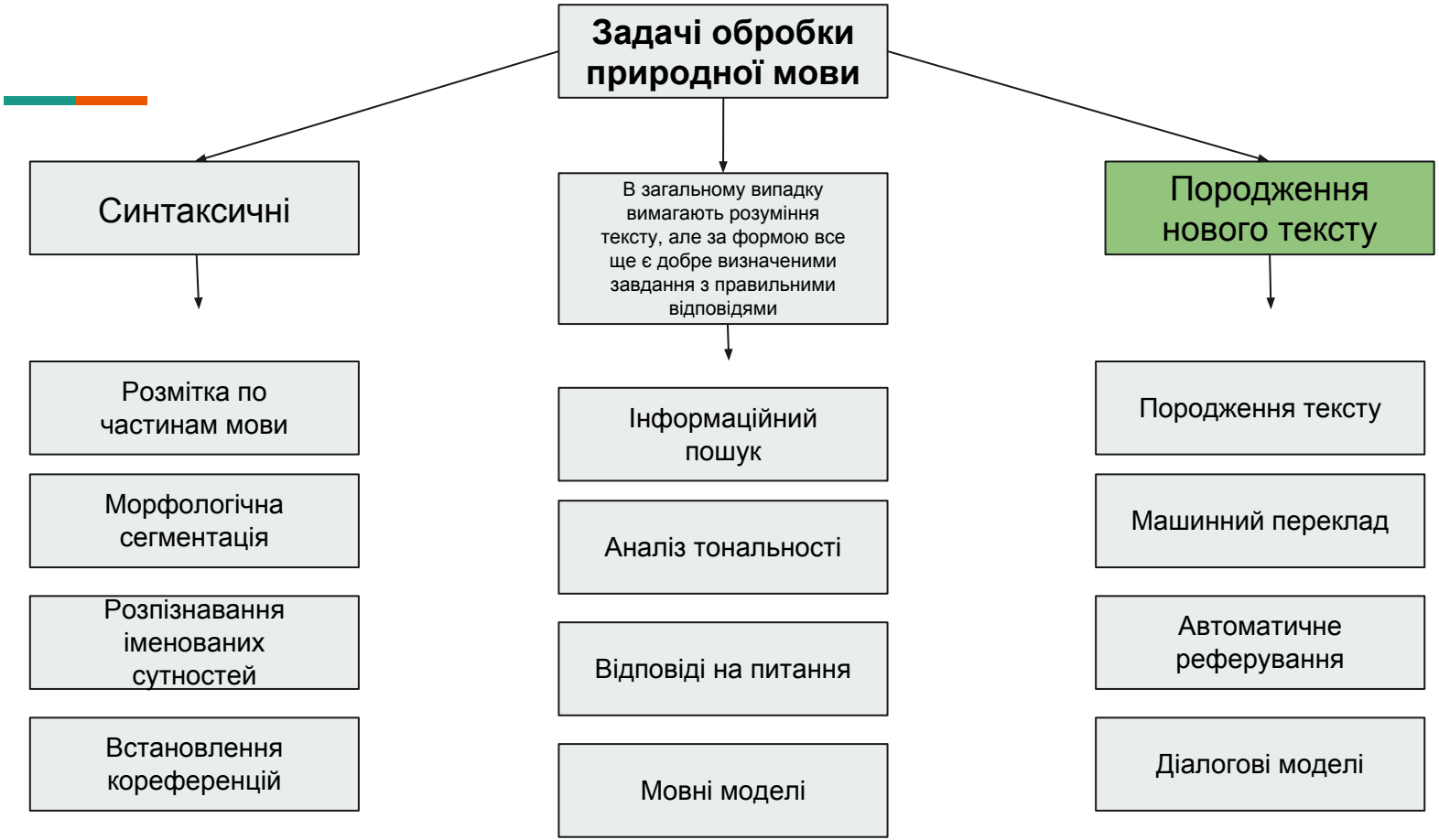
- автореферування статей та книжок;
- автоматична анотація зображень;
- автоматичний опис відео;
- автоматичне створення опису товару чи послуги з набору ключових слів;
- заповнення сайтів унікальним (не обов’язково осмисленим) контентом тощо.

Обробка природних мов

Обробка природних мов – це область проектування методів та алгоритмів, які використовують як вхідний матеріал або видають як вихід дані, що не є структурованими, природну мову.

Людська мова

- неоднозначна;
- мінлива;
- символьна;
- дискретна;
- композиційна.


Задачі обробки природної мови
Синтаксичні
В загальному випадку вимагають розуміння тексту, але за формою все ще є добре визначеними завдання з правильними відповідями
Породження нового тексту
Розмітка по частинам мови
Морфологічна сегментація
Розпізнавання іменованих сутностей
Встановлення кореференцій
Інформаційний пошук
Аналіз тональності
Відповіді на питання
Мовні моделі
Породження тексту
Машинний переклад
Автоматичне реферування
Діалогові моделі

Рекурентні нейронні мережі

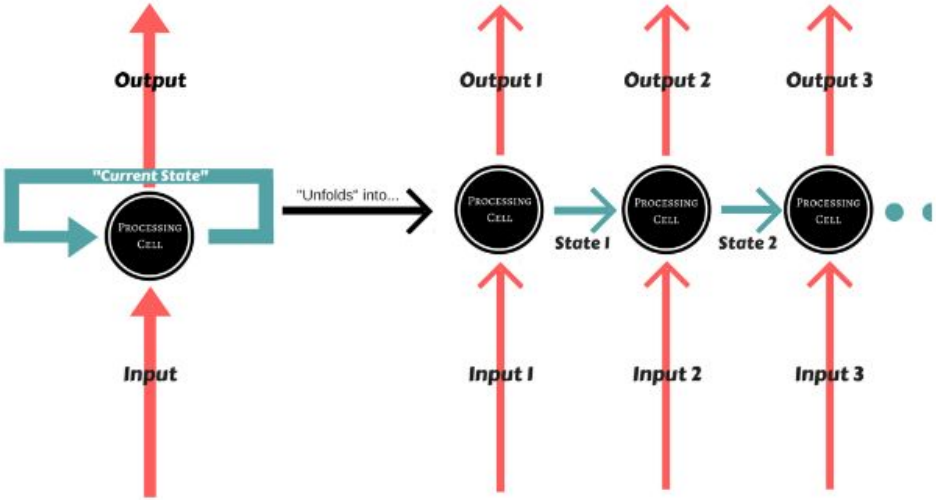


Схема LSTM комірки

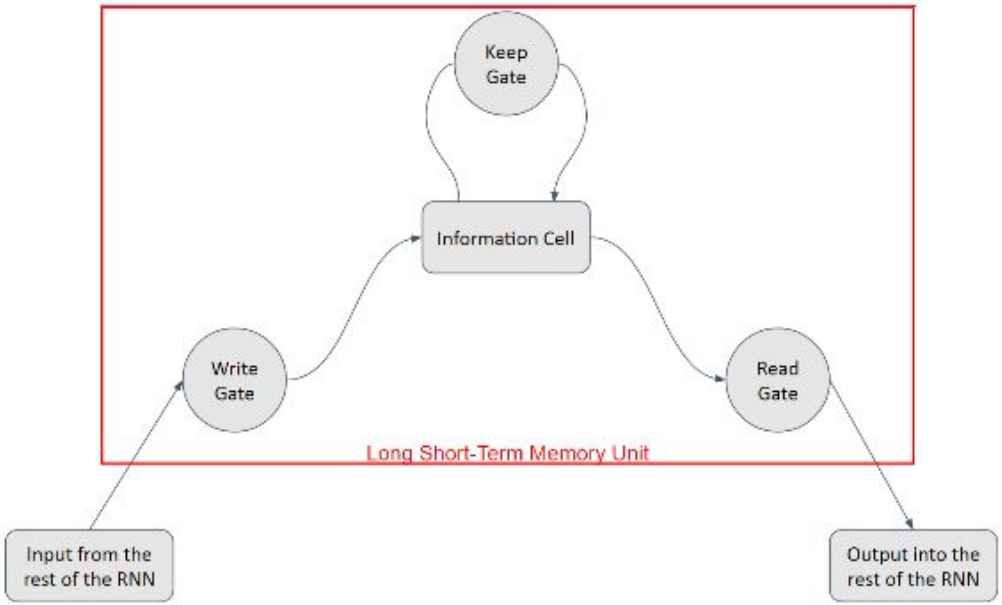
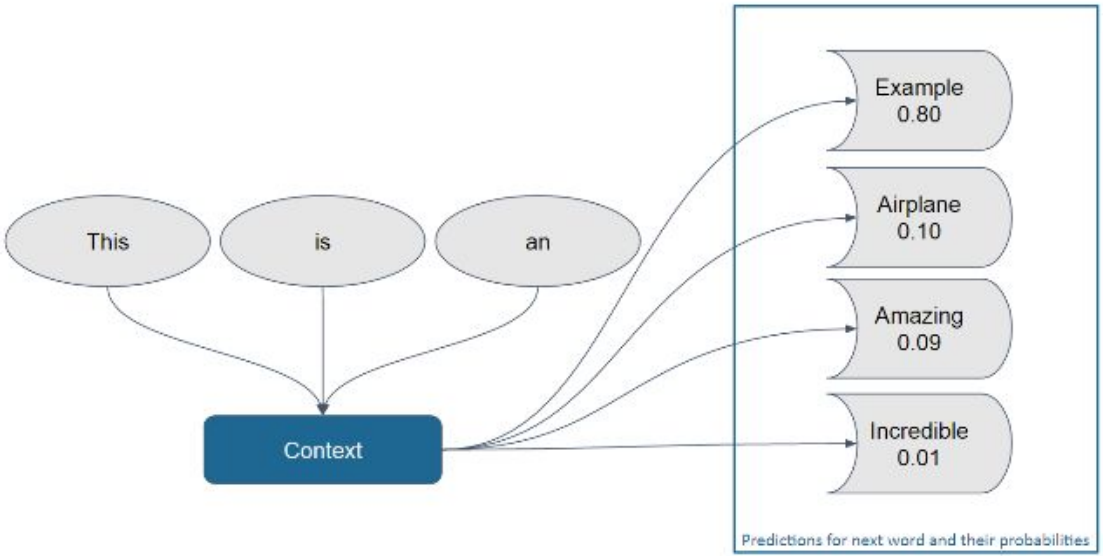
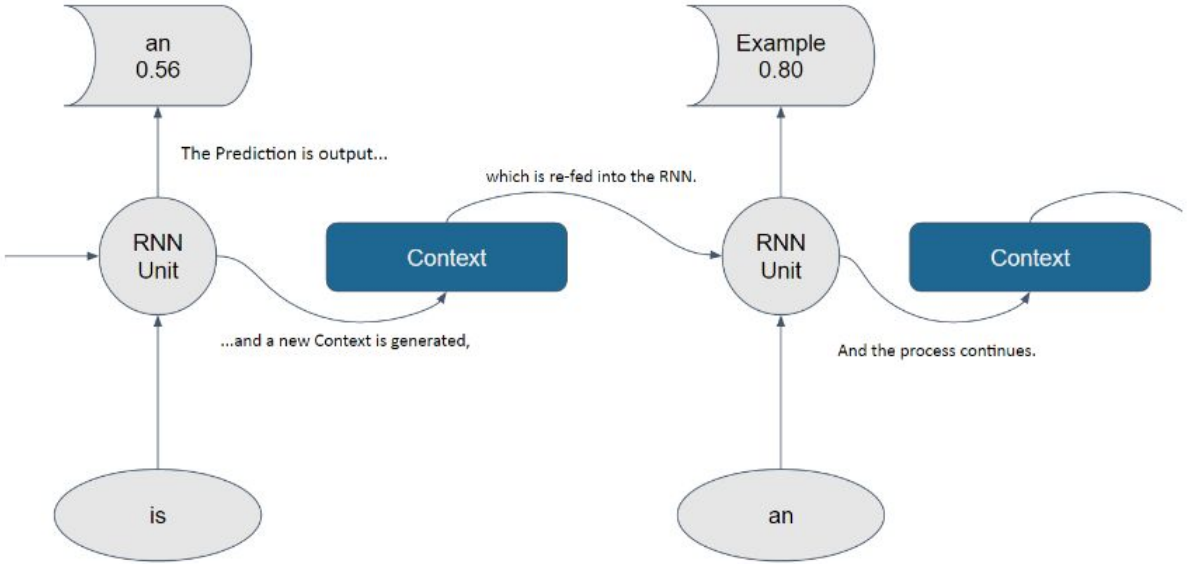


Схема генерування тексту мережею



Example of a sentence being predicted

Схема генерування тексту мережею

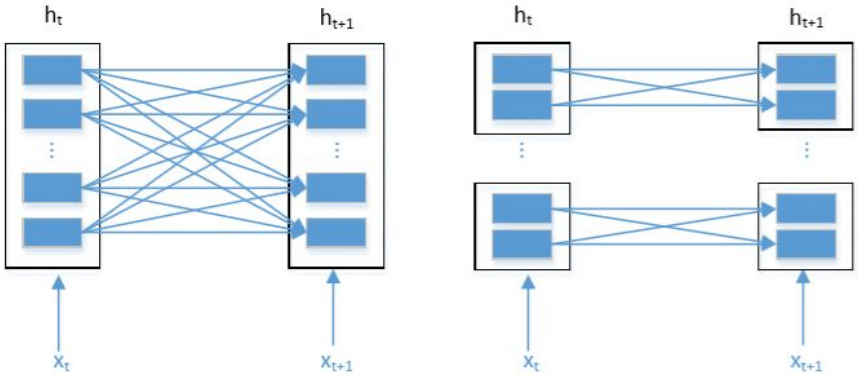


The above example schematized as an RNN in execution

Метрика якості моделі

Кращою на даний момент для подібних задач є оцінка заплутаності або невизначеності (perplexity) на небачених раніше реченнях. Заплутаність - це інформаційне теоретичне вимірювання того, наскільки добре вибрана модель прогнозує вибірку. Низькі значення свідчать про вищу впевненість моделі в прогнозі.

Розділення комірок на декілька відеокарт



Приклад згенерованого тексту

KING LEAR

Marry so and to the ground have yours

And thou Juliet my sir?

JULIET

Ah wherefore speak your satisfaction brother

May thee flee.

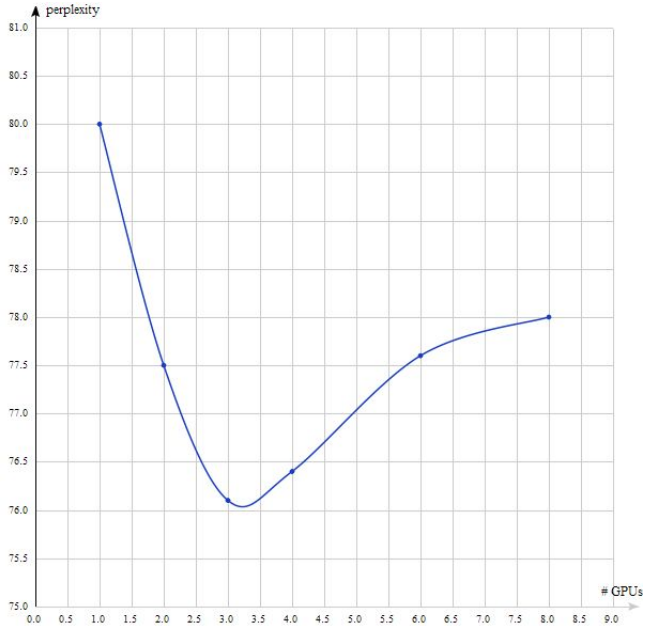
Приклад згенерованого тексту

Мы старались вывезти свою летнюю возлюбленную и у не одна. Кейт выглядит лучше грубые и оригинальные модели которых вы можете сделать фильм по крайней мере у нее богатые режиссеры весело или подчёркивает идеи в кино и лишь сцену картины. . . .. О спасибо всем наших многочисленных актрис.

Якість моделі для датасету Шекспіра для різної кількості відеокарт після 40 епох тренування

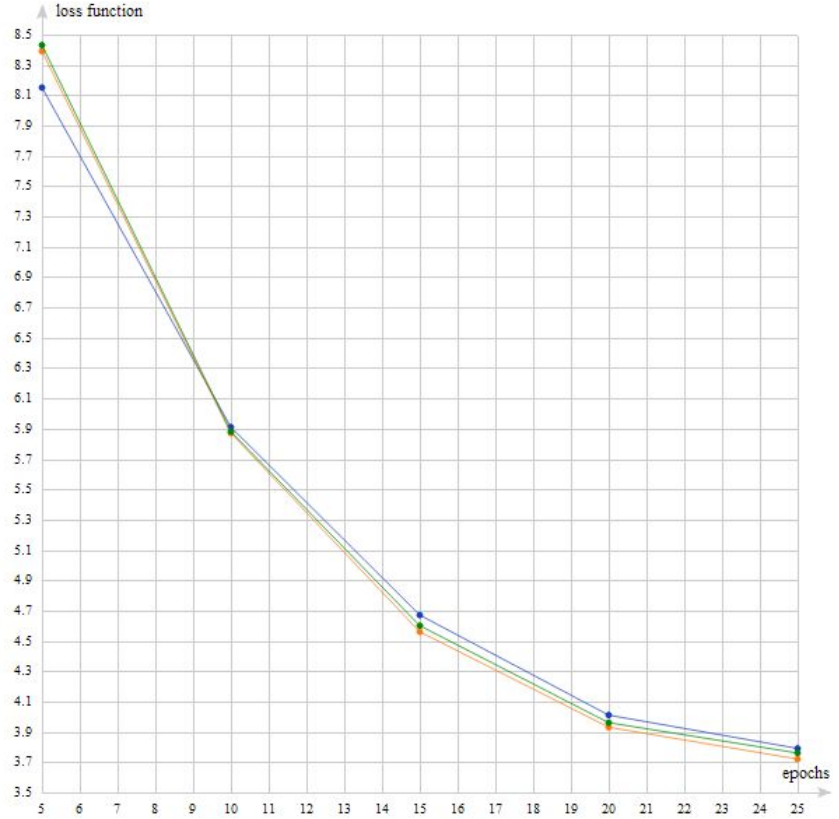
Кількість відеокарт (GPU)	Ширина шару	Розмір моделі, МБайт	Perplexity
1	512	57	80.0
2	256	28.7	77.5
3	172	19.1	76.1
4	128	14.5	76.4
6	84	9.7	77.6
8	64	7.2	78.0

Якість моделі для датасету Шекспіра для різної кількості відеокарт після 40 епох тренування



Швидкість навчання з використанням різної кількості відеокарт

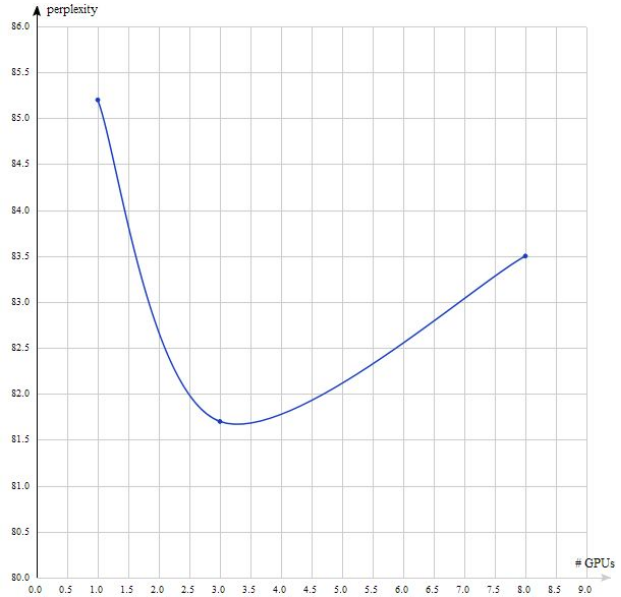
- синім - 1 GPU
- жовтим - 3 GPUs
- зеленим - 6 GPUs



Якість моделі для датасету з кінотематикою для різної кількості відеокарт після 20 епох тренування

Кількість відеокарт (GPU)	Ширина шару	Perplexity
1	512	85.2
3	172	81.7
8	64	83.5

Якість моделі для датасету з кінотематикою для різної кількості відеокарт після 20 епох тренування



Порівняння з іншими моделями, навченими на датасеті Шекспіра

Модель	Perplexity
Моя модель (RNN з 2 прихованими шарами, на 3 відеокатах, ширина шару по 172 на кожній відеокарті)	76.1
LSTM-1500h(Zaremba et al., 2014)	77.4
LSTM-Char(Kim et al, 2016)	77.9

Висновки

Реалізована система дозволяє розв’язати реальну бізнес-задачу. Запропонована модифікація зменшує розмір результуючої моделі, що дозволяє зекономити на носіях інформації, та має трохи вищу якість, ніж аналоги. Вона підходить як для латиниці, так і для кирилиці. Оптимальним варіантом виявивилось навчати модель на 3-4 відеокартах, що зменшило метрику perplexity.

Дякую за увагу!