Морфологічна розмітка

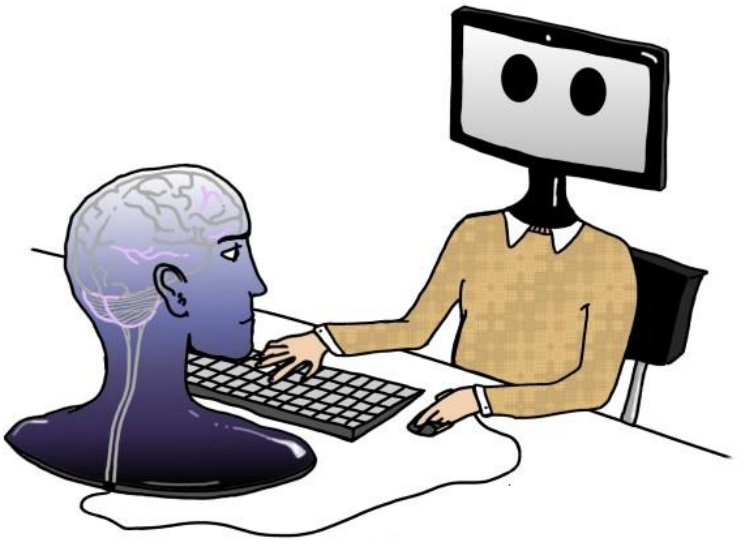
Чернівецький національний університет
імені Юрія Федьковича

Наукова студентська конференція
2016 рік

Доповідач: Дубінін Данило
Науковий керівник: доц. Сопронюк Т.М.

Обробка природних мов

- Загальний напрямок
  - Інформатики
  - Штучного інтелекту
  - Математичної лінгвістики



- Для взаємодії
  - Людина ↔ Машина ↔ Людина

Завдання

- Машинний переклад
- Розпізнавання мови
- Видобування даних (аналіз тексту)
- Синтез мовлення
- Інформаційний пошук, витяг
- Реферування
- Розв’язання лексичної багатозначності

Приклад Wolfram|Alpha

```
WolframAlpha["How many vehicles are there?", ExcludePods → {"Input"}]
```

Summary:

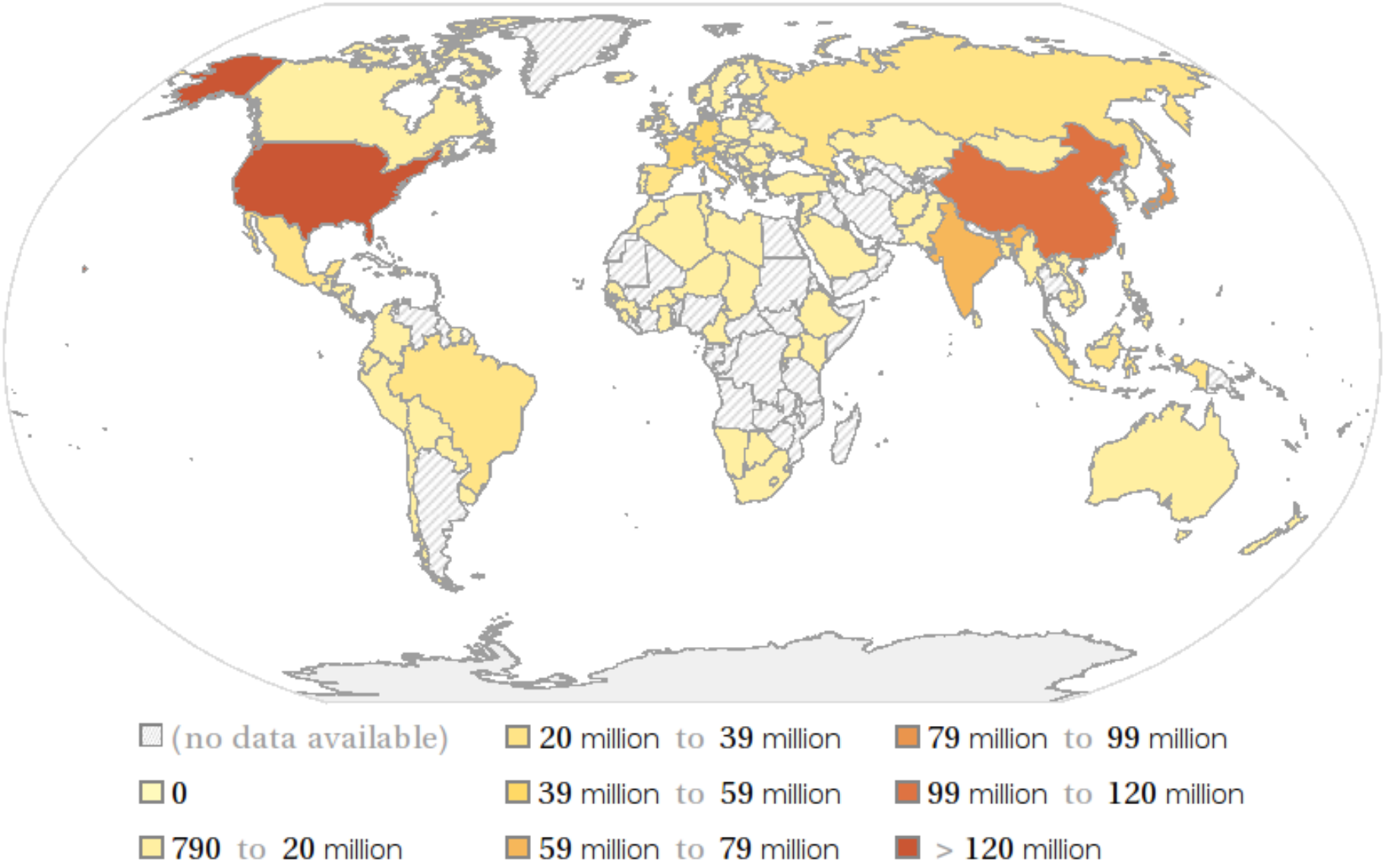
Show details

total	1.1 billion vehicles
median	1.1 million vehicles
highest	254.4 million vehicles (United States)
lowest	790 vehicles (Comoros)

(2001, 2002, 2003, 2004, 2005, 2006, 2007, and 2009 estimates)

(based on 123 values; 117 unavailable)

Приклад Wolfram|Alpha



Приклад Wolfram|Alpha

- **Make oxygen from CO2?**

CO_2 (carbon dioxide) + Na_2O_2 (sodium peroxide) $\longrightarrow$ O_2 (oxygen) + Na_2CO_3 (soda ash)

CO_2 (carbon dioxide) + KO_2 (potassium superoxide) $\longrightarrow$ O_2 (oxygen) + K_2CO_3 (pearl ash)

CO_2 (carbon dioxide) + H_2O (water) + KO_2 (potassium superoxide)
$\longrightarrow$ O_2 (oxygen) + $KHCO_3$ (potassium bicarbonate)

CO_2 (carbon dioxide) + H_2O_2 (hydrogen peroxide)
$\longrightarrow$ O_2 (oxygen) + H_2O (water) + CO (carbon monoxide)

CO_2 (carbon dioxide) + $HOOCCH_2CH_2COOH$ (succinic acid)
+ $C_{14}H_{19}N_3O_7S$ (Deacetylcephalosporin C) $\longrightarrow$
O_2 (oxygen) + $C_5H_6O_5$ (2–oxoglutaric acid) + $C_{14}H_{19}N_3O_6S$ (deacetoxycephalosporin–c)

Машинний переклад

- Методи
  - Правил
  - Ститистичні
  - Гібридні

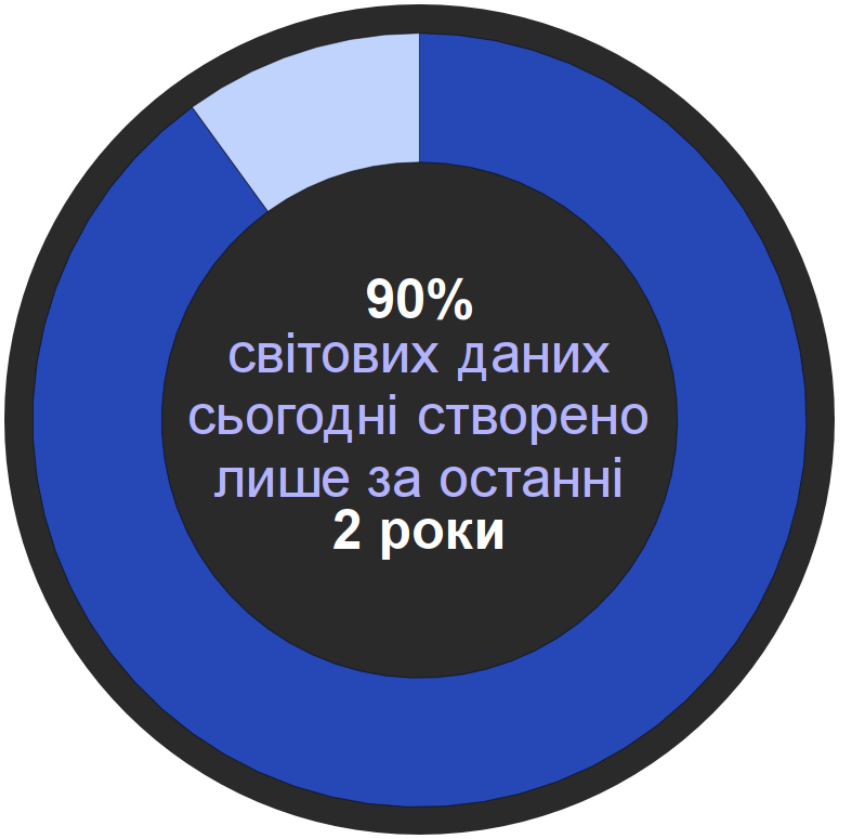
- Найскладніші проблеми
  - Лексична неоднозначність
  - Власні назви

Лексична багатозначність

Гриф

Видобування даних

- Виявлення закономірностей у великих необроблених масивах даних



Історія

- I (1940-1960)
  - Georgetown-IBM experiment (1954)
  - Russian → English
- II (1960-1970)
  - SHRDLU
- III (1970-1980)
  - Штучний інтелект
  - Логічне програмування
- IV (1990-)
  - Статистичні методи

I. Georgetown-IBM experiment

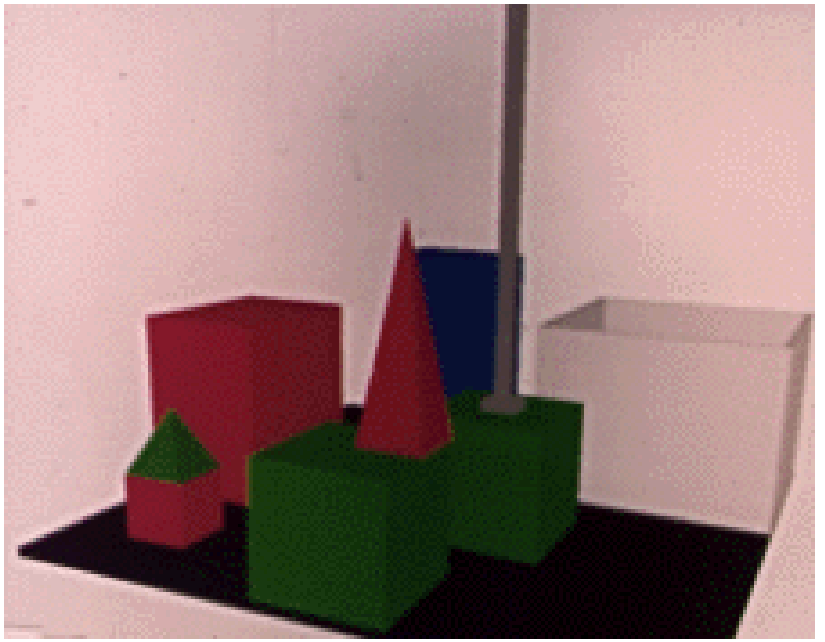
- Перший машинний переклад оснований на правилах
  - Перестановка слів
  - Вставка/вилучення слів
  - І т.п.

```
Mi pyeryedayem mislyi posryedstvom ryechyi.
```

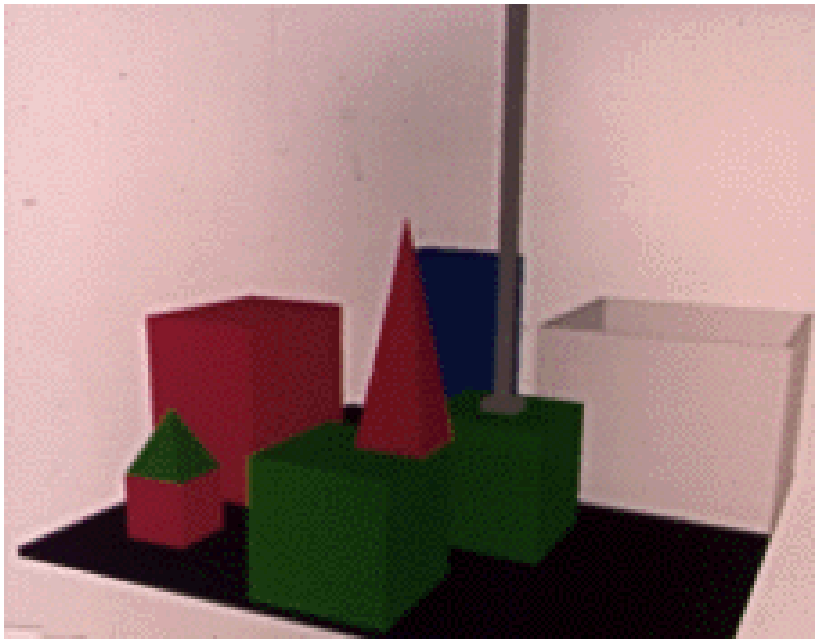
We transmit thoughts by means of speech.

II. SHRDLU

- P: Pick up a big red block.
- C: OK.
- P: Grasp the pyramid.
- C: I DON'T UNDERSTAND WHICH PYRAMID YOU MEAN.



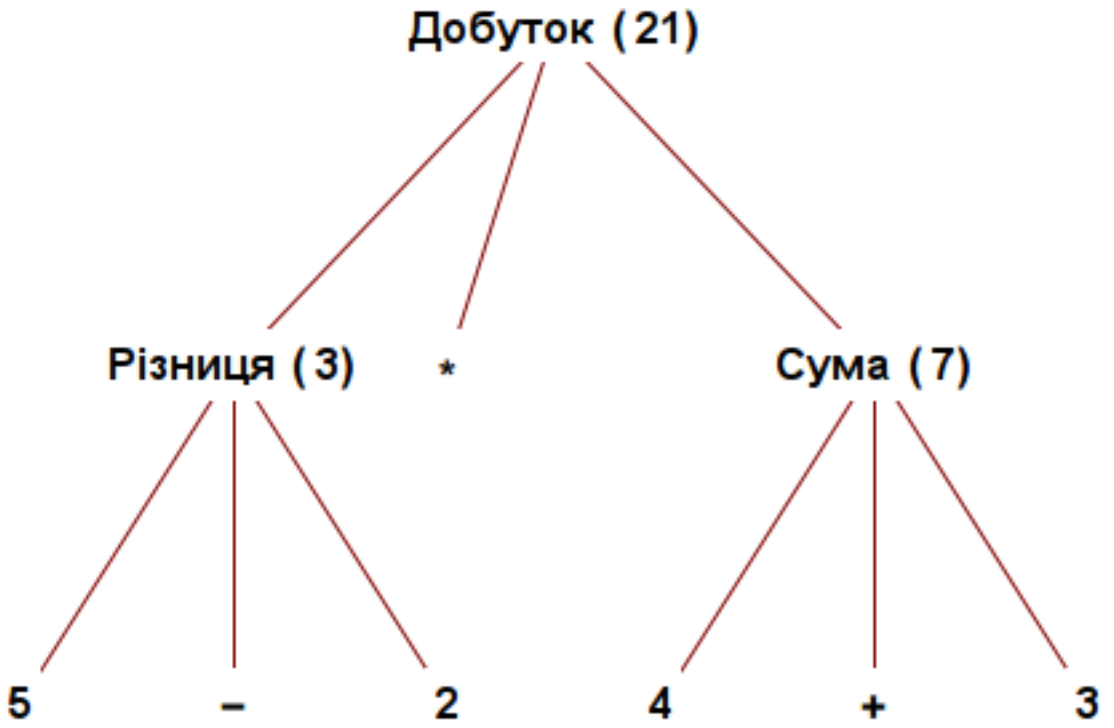
II. SHRDLU



- P: How many blocks are not in the box?
- C: FOUR OF THEM.
- P: Is at least one of them narrower than the one which i told you to pick up?
- C: YES, THE RED CUBE.

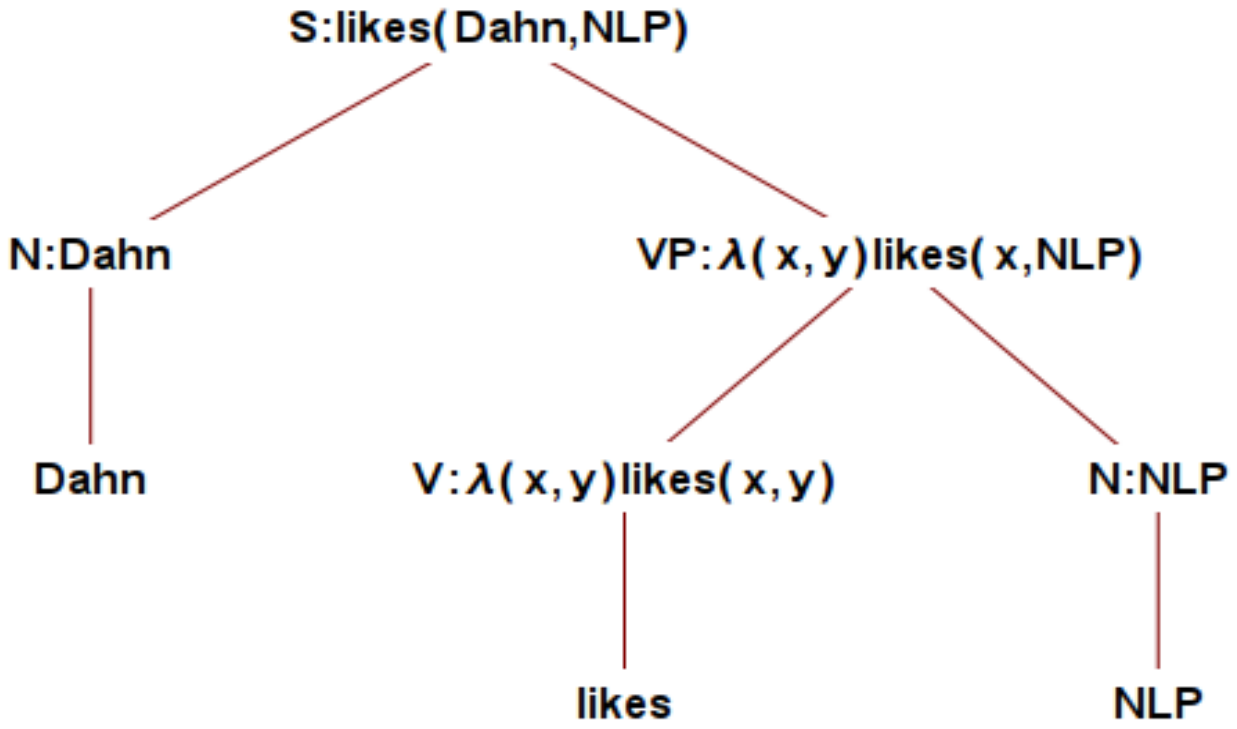
III. Синтаксичний розбір

- Матиматичні вирази? – Просто!
  - **(5-2)*(4+3)**



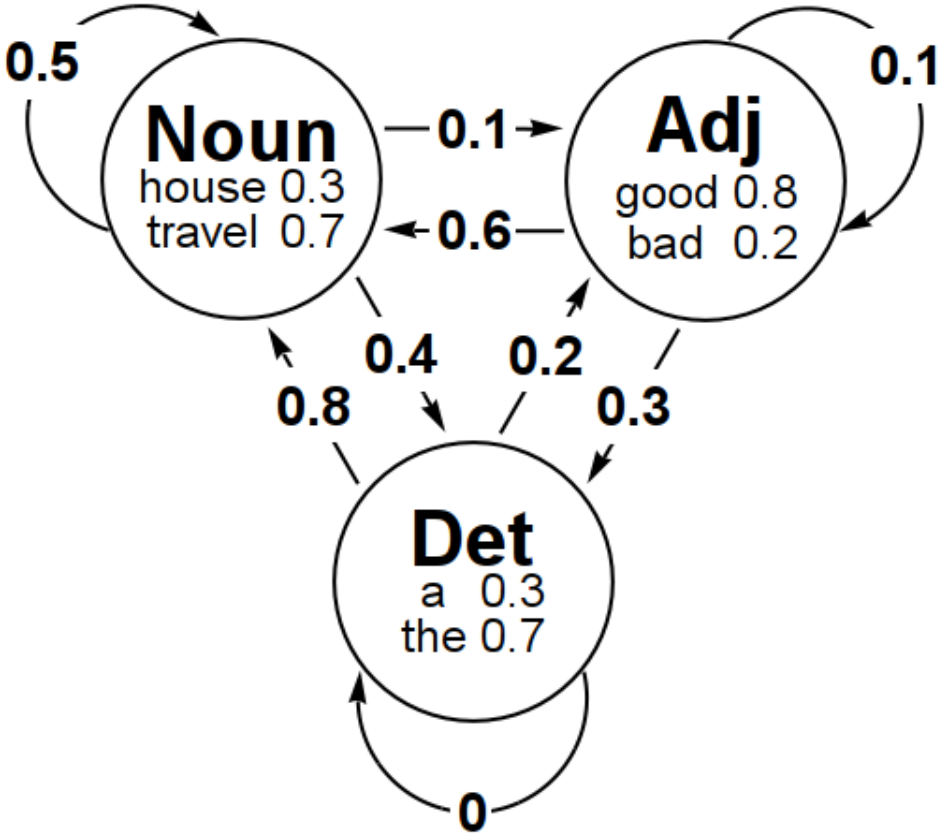
III. Синтаксичний розбір

- Речення природної мови?
- Предикативне числення!



IV. Статистичні методи

- Стохастичні контекстно-залежні граматики
- Прихована марковська модель (ППМ)



ШІ-повні задачі (AI-complete)

- Комп’ютерний зір
- Розуміння природної мови
- Розв’язування реальних проблем за непередбачуваних обставин

ШІ-повні задачі (AI-complete)

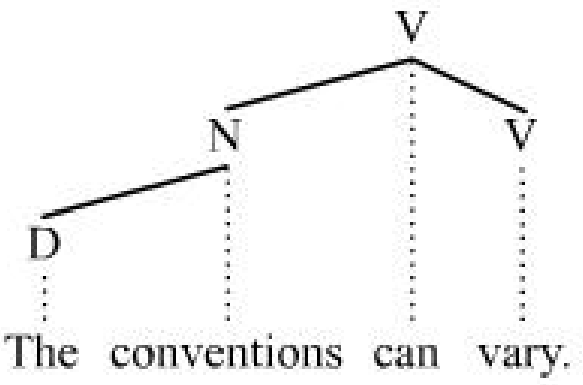
- Комп’ютерний зір
- Розуміння природної мови
- Розв’язування реальних проблем за непередбачуваних обставин (навігація, планування, експертні системи, і т.п.)

Морфологічна розмітка

- Дано речення: послідовність слів $\boldsymbol{x}$
- Визначення частини мови слів $\boldsymbol{y}$
  - ADJ, ADP, ADV, CONJ, DET, NOUN, NUM, PRON, PRT, VERB
  - . – пунктуація
  - <S>, </S> – початок/цінець речення
  - X – невідома частина мови

Навіщо

- Розв’язання лексичної неднозначності
  - Покращення результату використовуючи оточуючі частини мови як характеристики
- Переклад
  - Додаткове знання частин мови не заважає
- Синтаксичний розбір

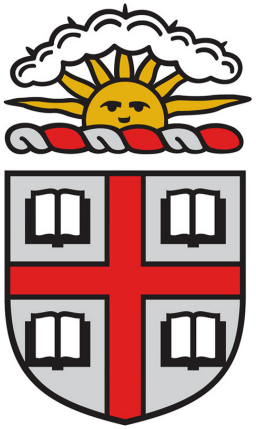


Схема

- Обчислення параметрів ППМ
  - Машинне навчання
- Алгоритм Вітербі
  - Динамічне програмування

Навчання

- *Лінгвістичний корпус*: Brown corpus
  - складений у 1960-х
  - 500 текстів
  - 1 млн. слів
  - людьми визначенні частини мови всіх слів
  - Once/ADP the/DET principle/NOUN was/VERB established/VERB ,/. the/DET increase/NOUN in/ADP state-owned/ADJ vehicles/NOUN came/VERB rapidly/ADV ./.



Прихована марковська модель

- ПММ – марковський процес із *неспостережуваними станами*.
- *Спостереження*
  - Слова речення
  - **The quick brown fox jumps over the lazy dog.**
- *Стани*
  - Частини мови
  - **DET ADJ ADJ NOUN VERB PREP DET ADJ NOUN**

ПММ морфологічної розмітки

- Дано ланцюжки пар: < слово, частина мови >
  - $(\mathrm{x})_i, (\mathrm{y})_i$
- $f : X \to Y$ **?**
- Вивчаємо розподіл $\mathbb{P}(\mathrm{y}|\mathrm{x})$
- Знаходимо розв’язок у вигляді

$$f(\mathrm{x}) = \mathrm{argmax}_y \mathbb{P}(\mathrm{y}|\mathrm{x})$$

ПММ морфологічної розмітки

$$y = \operatorname{argmax}_y$$

$$\underbrace{\prod_i \mathbb{P}(x_i \mid y_i)}_{\text{спостереження}} \times$$

Слова

$$\times \quad \underbrace{\prod_i \mathbb{P}(y_i \mid y_{i-1})}_{\text{стани}}$$

Частини мови

Алгоритм Вітербі

- Таблиця динамічного програмування

$$\pi_k(\mathrm{A},\mathrm{B}) = \max_{\substack{\langle y_1 \dots y_k \rangle : \\ y_{k-1}=A,\, y_k=B}} \prod_{i=1}^{k} \mathbb{P}(x_i \mid y_i)\mathbb{P}(y_i \mid y_{i-1})$$

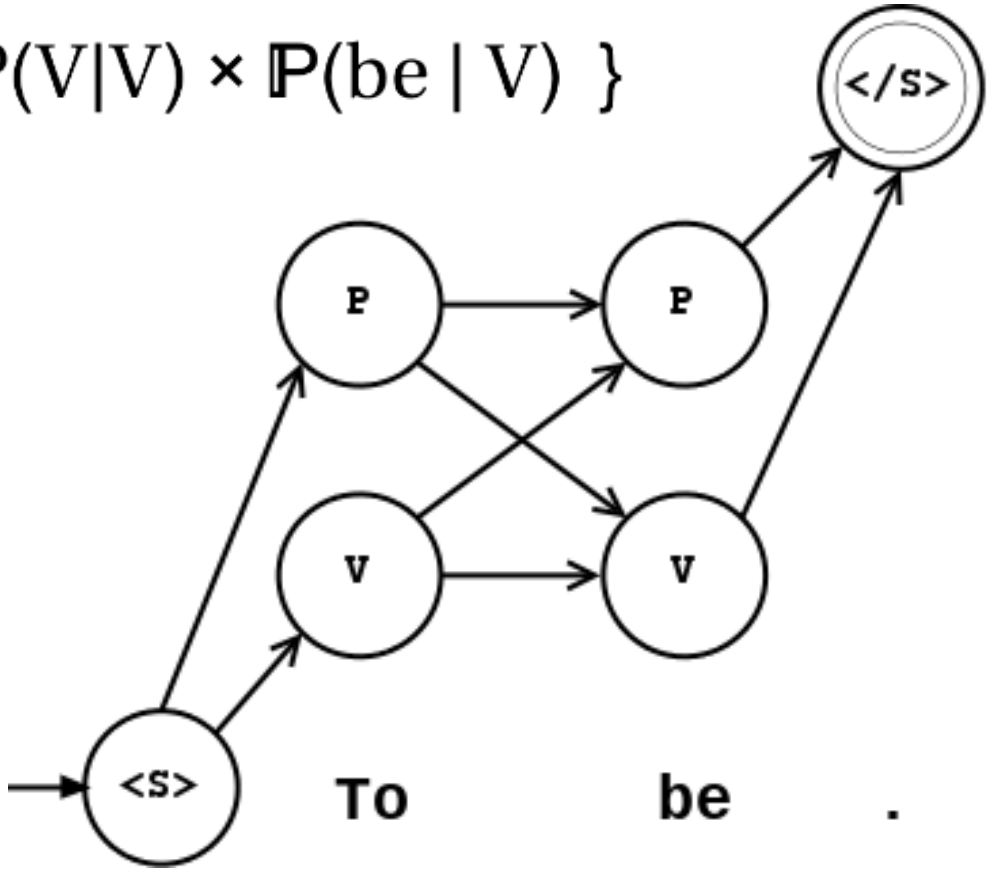
- Рекурсивне обчислення

$$\pi_k(\mathrm{Y},\mathrm{Z}) = \max_X (\quad \pi_{k-1}(\mathrm{X},\mathrm{Y}) \times \mathbb{P}(\mathrm{Z} \mid \mathrm{X},\mathrm{Y}) \times \mathbb{P}(x_k \mid \mathrm{Z}) \quad)$$

$$\pi_0 = 1$$

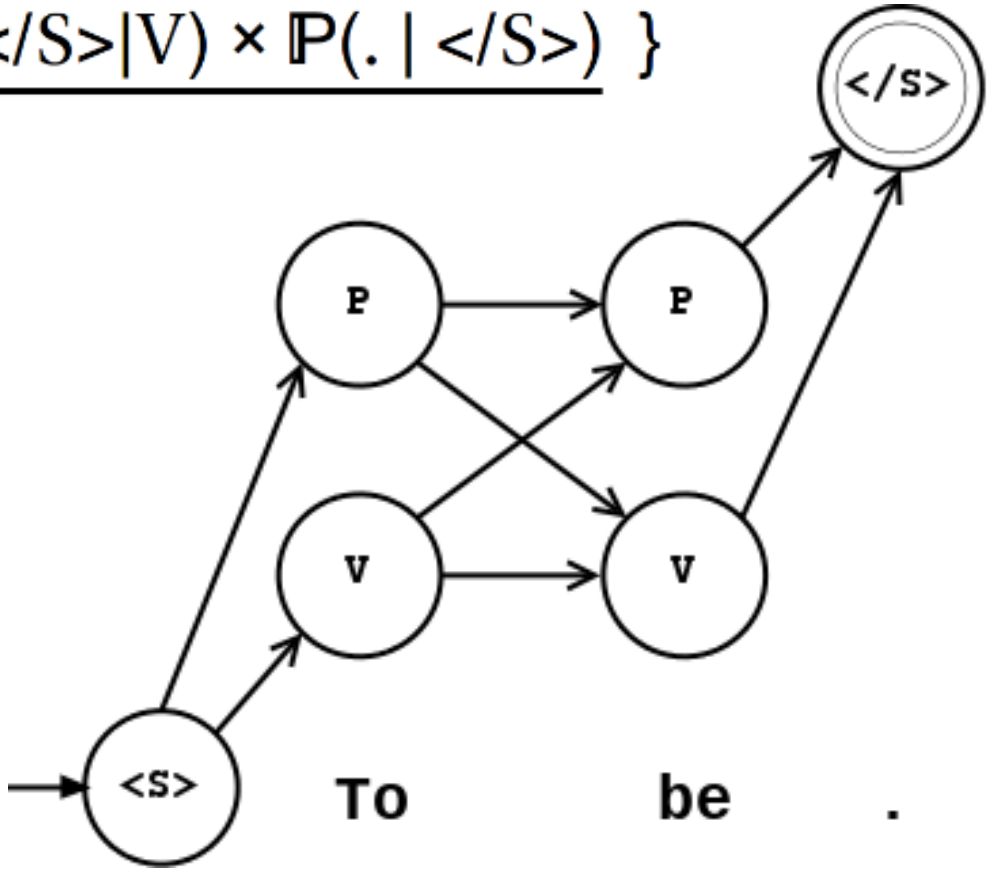
Приклад

$$\pi_2(\mathrm{V}) = \max\{ \frac{\pi_1(\mathrm{P}) \times \mathbb{P}(\mathrm{V}|\mathrm{P}) \times \mathbb{P}(\mathrm{be} \mid \mathrm{V})}{\pi_1(\mathrm{V}) \times \mathbb{P}(\mathrm{V}|\mathrm{V}) \times \mathbb{P}(\mathrm{be} \mid \mathrm{V})} \}$$



Приклад

$$\pi_3(\text{</S>}) = \max \{$$
$$\pi_2(\text{P}) \times \mathbb{P}(\text{</S>}|\text{P}) \times \mathbb{P}(. \mid \text{</S>}),$$
$$\underline{\pi_2(\text{V}) \times \mathbb{P}(\text{</S>}|\text{V}) \times \mathbb{P}(. \mid \text{</S>})} \}$$



Точність

- Власна реалізація
  - ~ 93%

- Межа професійних реалізацій
  - ~ 97% (2014 рік)

- Лінгвісти
  - ~ 98%

У дії

Sentence > I've given my great presentation for scientific student conference in CHNU.

Tagged : I/PRON 've/VERB given/VERB my/DET great/ADJ presentation/NOUN for/ADP scientific/ADJ student/NOUN conference/NOUN in/ADP CHNU/NOUN ./.

У дії

Sentence > I love those exciting travels we had last year!

Tagged : I/PRON love/VERB those/DET exciting/ADJ travels/NOUN we/PRON had/VERB last/ADJ year/NOUN !/.

У дії

Sentence > My best friend travels a lot.

Tagged : My/DET best/ADJ friend/NOUN travels/VERB a/DET lot/NOUN ./.

Але статистика не завжди права

чорний кіт тікає від білої собаки Translate

Ukrainian > English

black cat runs away from a white dog

чорний кіт тікає від білого собаки Translate

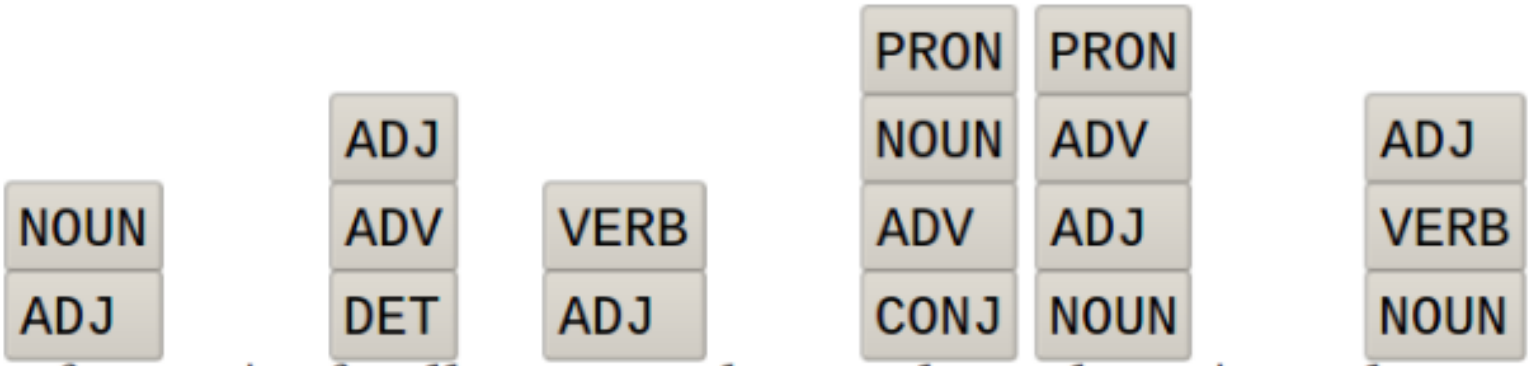
Ukrainian > English

black cat runs away from white dogs

Висновок

- Математичний підхід до лінгвістики!
- Прихована марковська модель ефективна
  - Для легких підзадач
- Для складніших задач питання відкриті
  - Поєднання математики та лінгвістики?
  - Використання більших навчальних баз та обчислюваних можливостей?

Тестування?



Perfection is *finally attained* not *when there* is no *longer* anything to add, but when there is no longer anything to take away.

Дякую :)

Perfection/NOUN is/VERB finally/ADV attained/VERB not/ADV when/ADV there/PRT is/VERB no/DET longer/ADJ anything/NOUN to/PRT add/VERB ,/. but/CONJ when/ADV there/PRT is/VERB no/DET longer/ADJ anything/NOUN to/PRT take/VERB away/ADV ./.