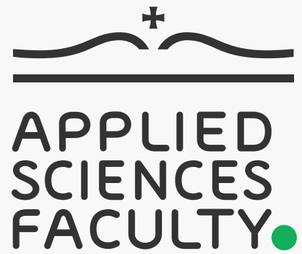
Nataliia Romanyshyn
Ukrainian Catholic University
romanyshyn.n@ucu.edu.ua

DMYTRO CHAPLYNSKYI
lang-uk
chaplinsky.dmitry@gmail.com

MARIANA ROMANYSHYN
Grammarly
mariana.romanyshyn@grammarly.com

Automated Extraction of Hypo-Hypernym Relations for the Ukrainian WordNet



I Research Background & Motivation

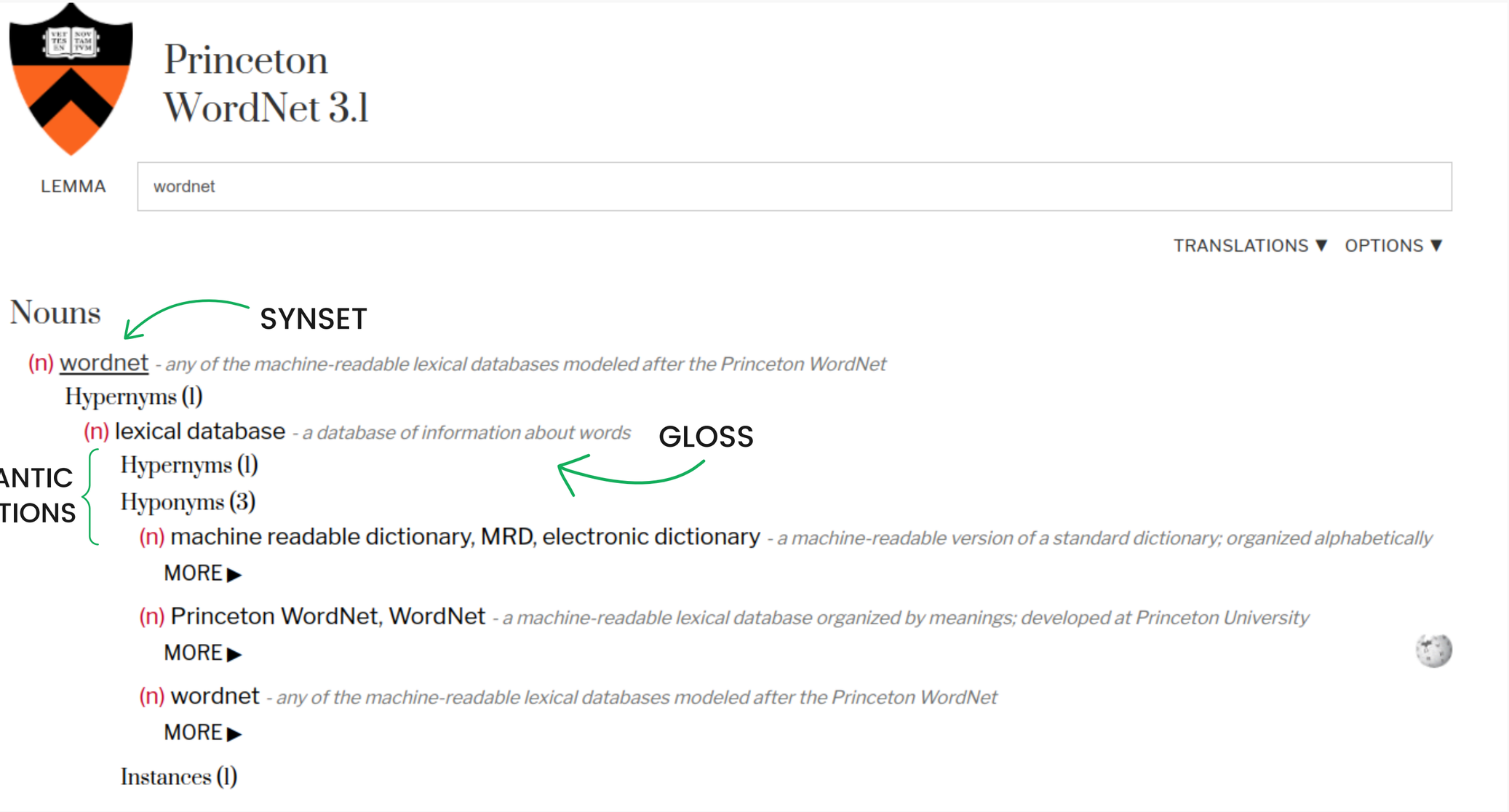
WordNet is a lexical database of semantic relations between words in a language that can be applied in various natural language processing and understanding tasks

The pioneer is the Princeton WordNet (PWN) of the English language (1994)

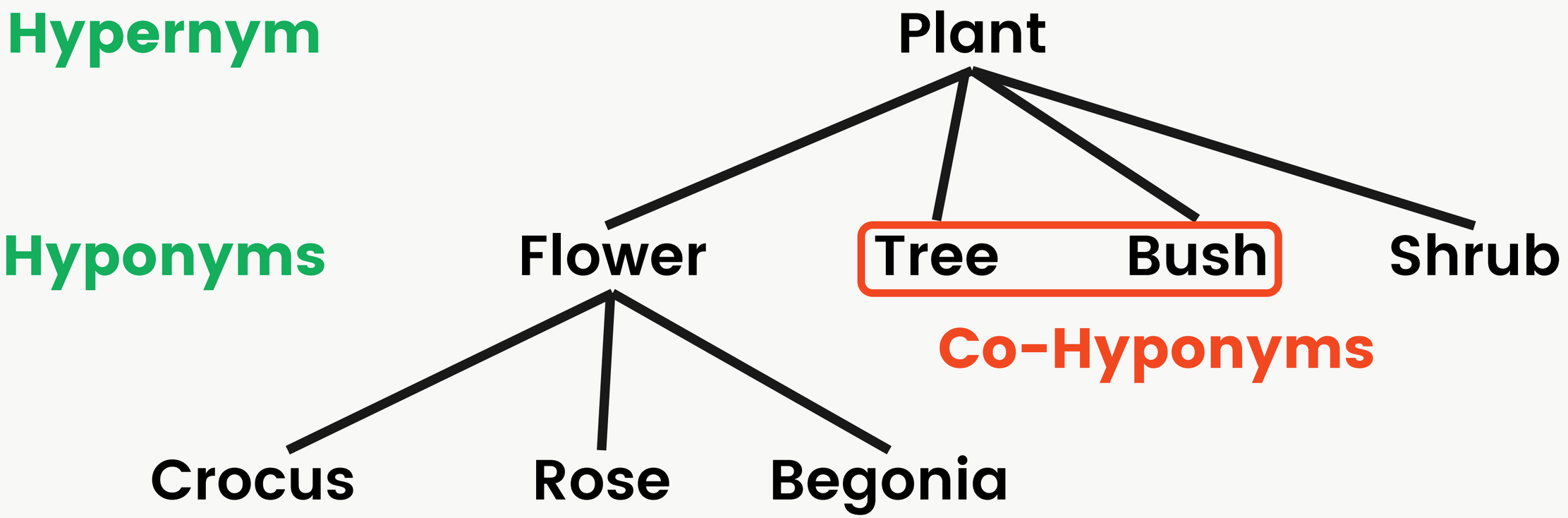
Automatic approaches for constructing and expanding WordNets have gained interest due to the high cost of manual taxonomy creation

There are now WordNets in more than 200 languages, but Ukrainian has yet to have a publically available one.

I Research Background & Motivation



Source: https://commons.wikimedia.org/wiki/File:Hyponym_and_hypernym.svg



An example of the hypernym, hyponyms, and co-hyponyms hierarchy.

Project Goals

01 Introduce a novel technique for creating a basis for Ukrainian WordNet

02 Develop an algorithm that maps Ukrainian Wikipedia titles to synsets in the Princeton WordNet

03 Develop a method of prioritizing the gap nodes in the Ukrainian WordNet

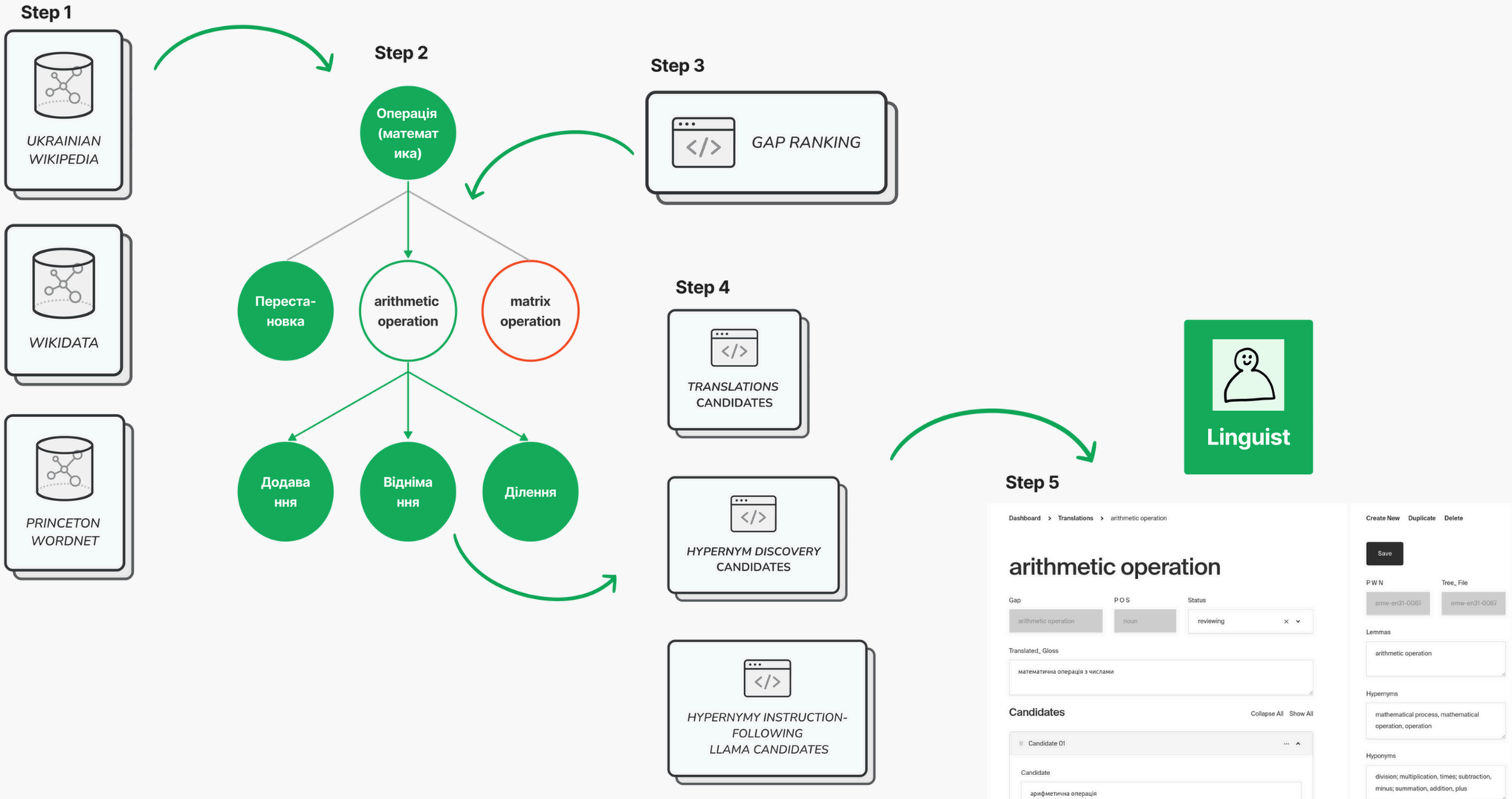
04 Propose strategies for automated generation of candidate words to fill the gaps

2009	NU "LP" MS Thesis by Khariv
2010	“Developing a WordNet-like Dictionary of Ukrainian” by Kulchytsky, **Romaniuk**, and Khariv
2013	“Ukrainian WordNet: Creation and Filling” by Anisimov, Marchenko, Nikonenko, Porkhun and Taranukha
2015	NU "LP" MS Thesis by Skopyk
2023	“Towards UkrainianWordNet: Incorporation of an Existing Thesaurus in the Domain of Physics” by Siegel, Vakulenko and Baum

Ukrainian WordNet: Status and Challenges

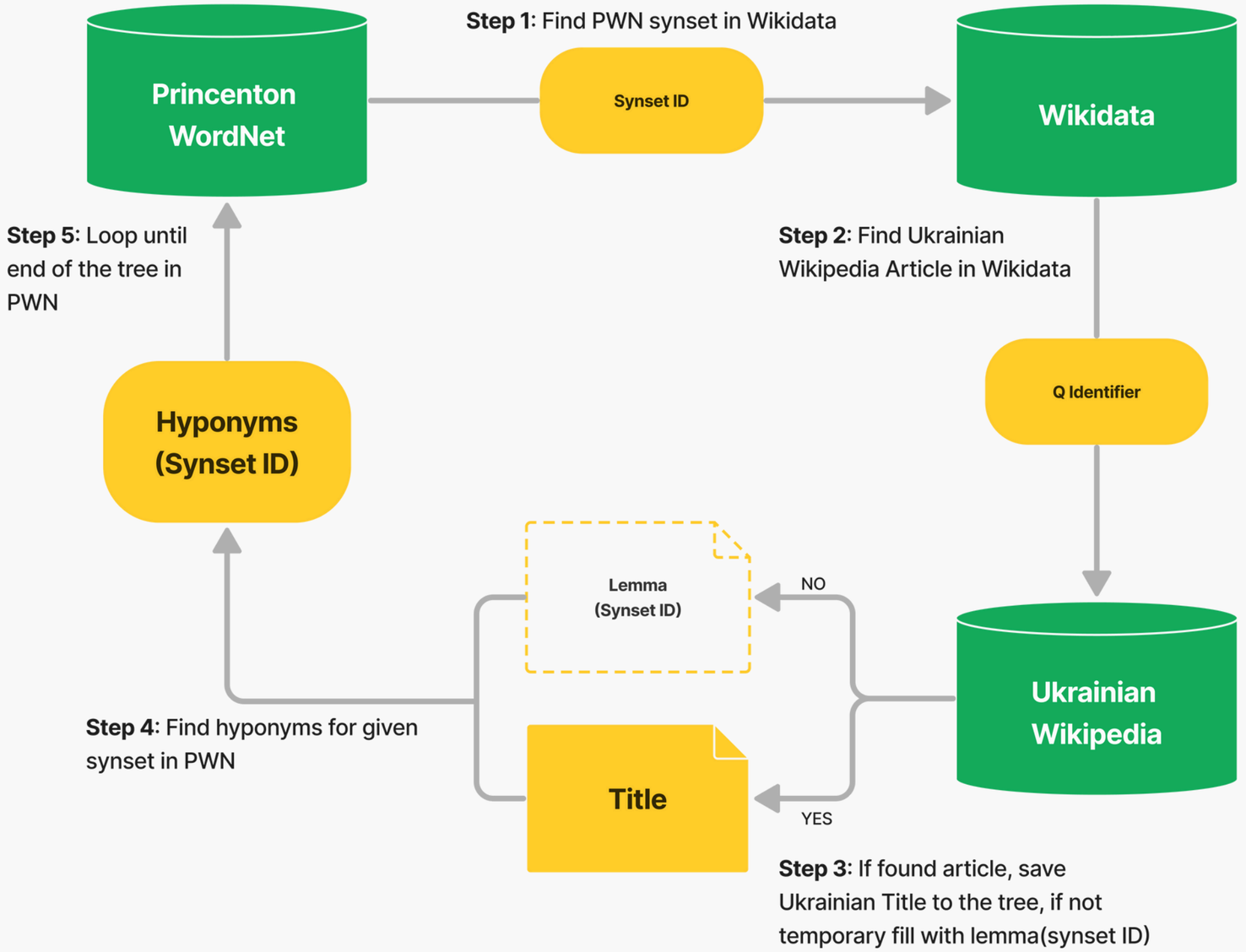
- The Ukrainian WordNet construction began in the 2010s with manual analysis and resulted in a WordNet-like dictionary with 194 synsets
- An automated approach developed UkrWordNet by generating nodes from Ukrainian Wikipedia articles, resulting in over 82,000 synsets
- Ukrajinet 1.0 centered around 3,360 synonym sets of physics terminology. However, it does not include hypo-hypernym relations
- Hence, developing an open-source WordNet for the Ukrainian language, with a representative number of relations, remains an ongoing area for research

III Methodology: proposed approach

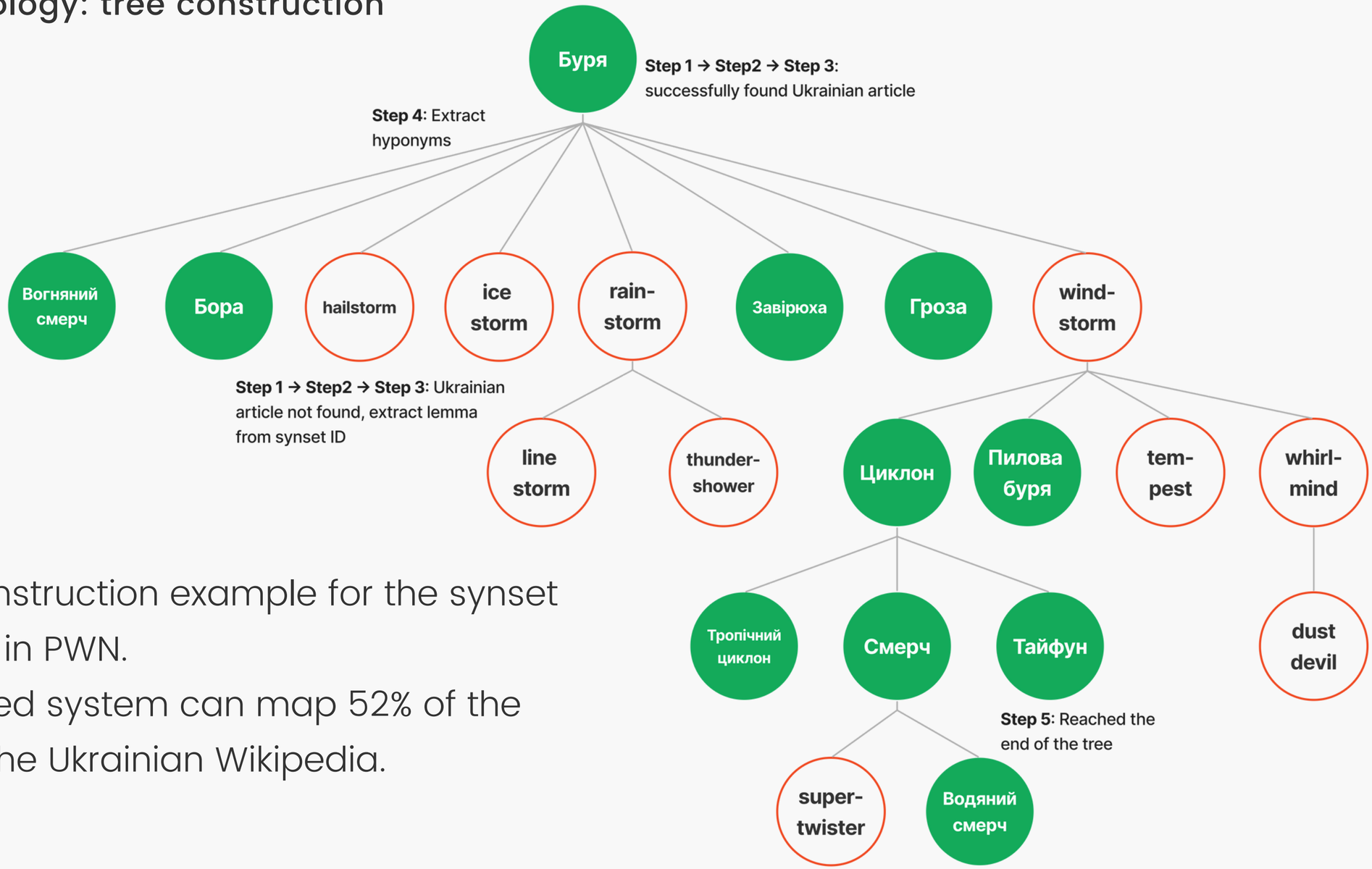


III Methodology: PWN, Wikidata and Ukrainian Wiki

Pipeline for building the basis of the Ukrainian WordNet utilizing the linking between Princeton WordNet, Wikidata, and the Ukrainian Wikipedia.



III Methodology: tree construction



The tree construction example for the synset 11482925-n in PWN.

The proposed system can map 52% of the subtree to the Ukrainian Wikipedia.

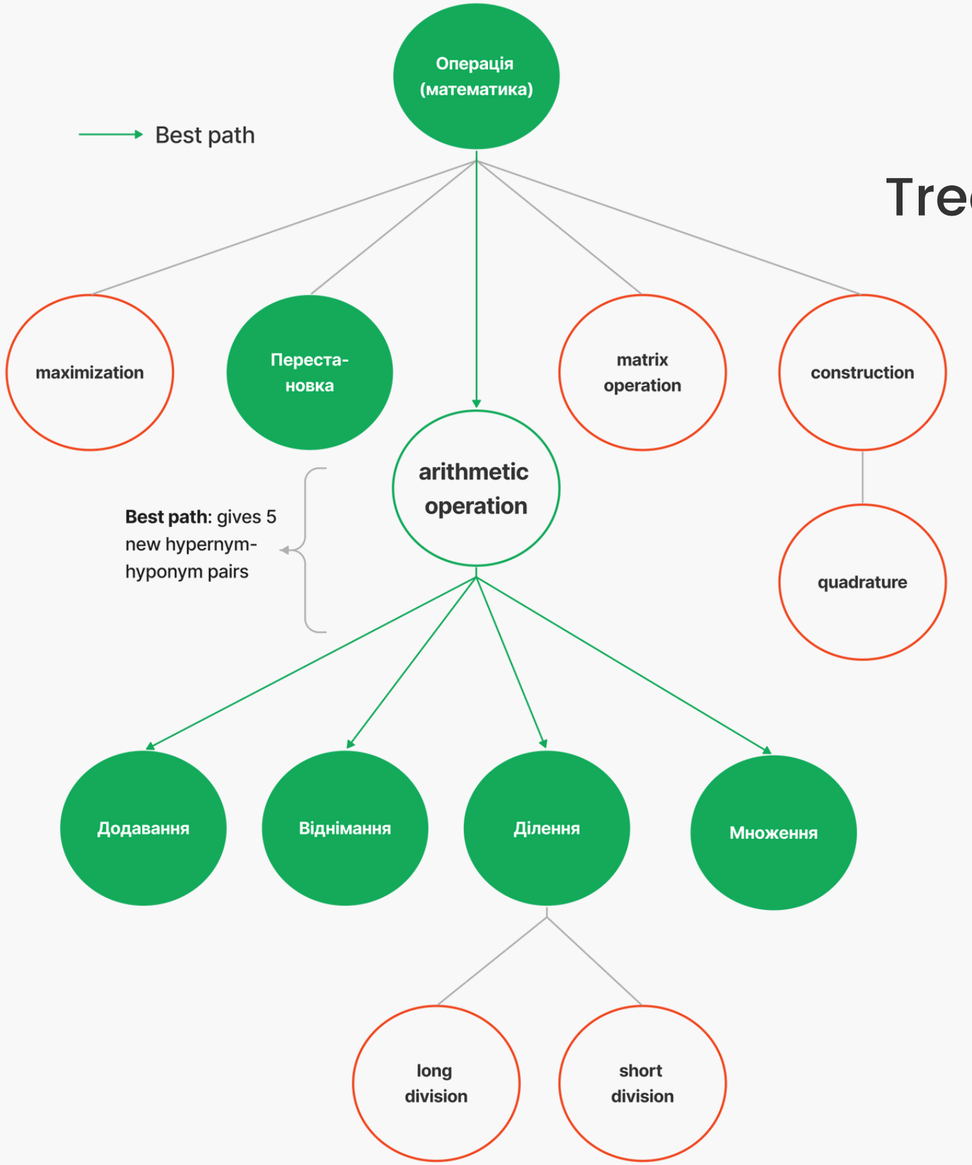
	# of synsets	% of synsets
PWN3.1	127,020	100%
Linked to Wikidata	29,730	23%
Linked to Ukrainian Wiki	21,015	17%

Linked WordNet basis statistics

Challenges

- synset ID is not linked with Wikidata
- the lack of a Ukrainian page on the wiki for the corresponding Wikidata Q identifier
- synset ID in Wikidata leads to an empty page

III Methodology: Gap Ranking

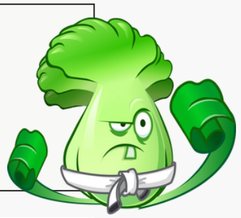


Tree fragment for synset **00871261-n**

Filling the node **arithmetic operation** is more effective than the other nodes, as it produces **five** new hypernym-hyponym pairs compared to only one pair for other nodes.

III Methodology: gap filling candidates generation

	DeepL Direct	DeepL Contextualized	Translated PWN3.1
performance	продуктивність **produktyvnist**	вистава **vystava**	вистава, спектакль **vystava, spektakl**
head cabbage	качанна капуста **kachanna kapusta**	качанна капуста **kachanna kapusta**	головна капуста **holovna kapusta**
agency	агентство **ahentstvo**	агентство **ahentstvo**	офіс, орган **ofis, orhan**



Comparison examples of gap translations obtained using machine translation methods.
All terms are nouns.

Hypernym Discovery

- Ukrainian adaptation to SemEval-2018 Task 9
- Utilized supervised part of the model, proposed by task winners Bernier-Colborne and Barrière(2018): pretrained word embeddings + logistic regression classifier
- Experimented with different embedding techniques (word2vec & fasttext)using the 31GB UberText 2.0 corpus

III Methodology: gap filling candidates generation

Hypernym Discovery

- Ukrainian adaptation to SemEval-2018 Task 9
- Utilized supervised part of the model, proposed by task winners Bernier-Colborne and Barrière(2018): pretrained word embeddings + logistic regression classifier
- Experimented with different embedding techniques (word2vec & fasttext)using the 31GB UberText 2.0 corpus

Generative AI

- Hypernymy Instruction-Following LLaMA (Large Language Model Meta AI) 7B
- Transformer-based Large Language Model
- Used a parameter-efficient tuning technique LoRA (Low-Rank Adaptation)
- Fine-tuning hyperparameters were taken from UAlpaca
- Build diverse experimental setups: lean, full and multiple

III Methodology: gap filling candidates generation

Hypernym Discovery

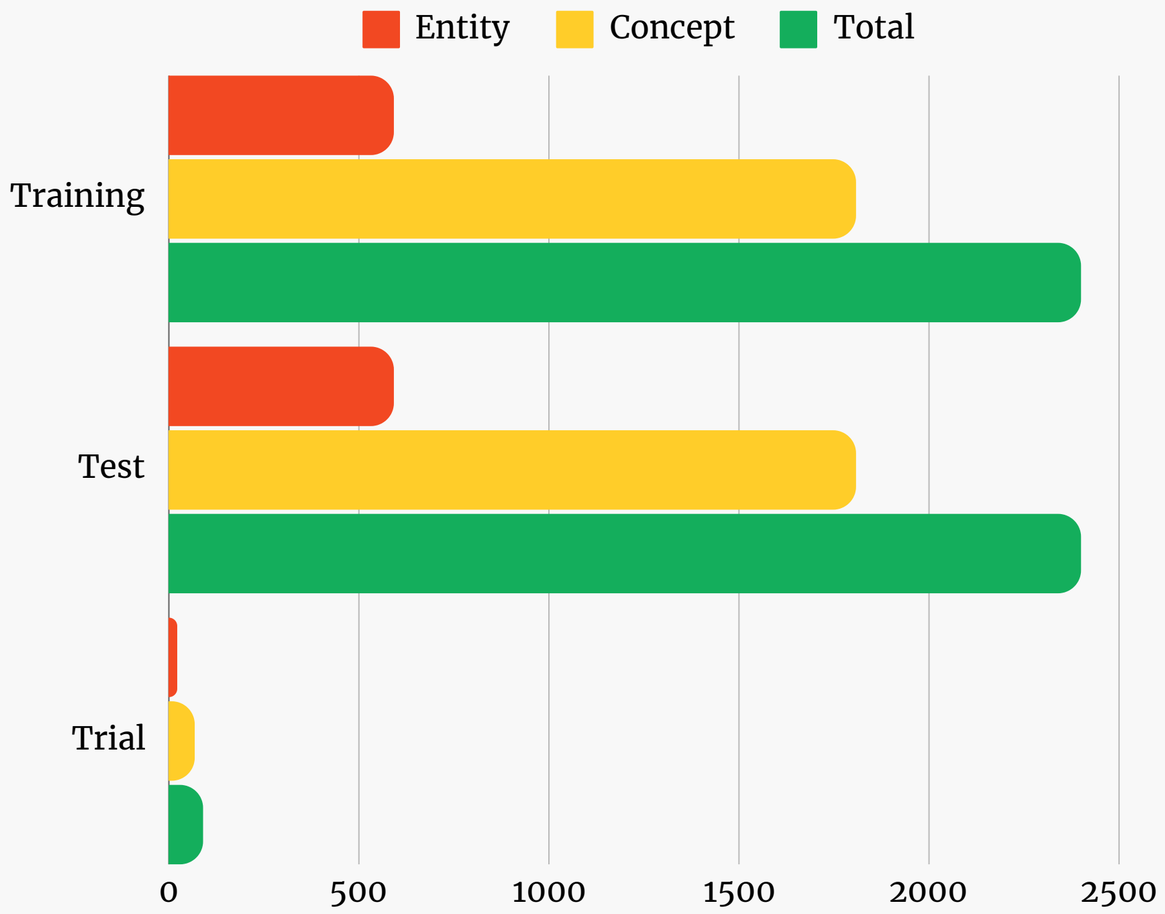
- Ukrainian adaptation to SemEval-2018 Task 9
- Utilized supervised part of the model, proposed by task winners Bernier-Colborne and Barrière(2018): pretrained word embeddings + logistic regression classifier
- Experimented with different embedding techniques (word2vec & fasttext)using the 31GB UberText 2.0 corpus

Generative AI

- Hypernymy Instruction-Following LLaMA (Large Language Model Meta AI) 7B
- Transformer-based Large Language Model
- Used a parameter-efficient tuning technique LoRA (Low-Rank Adaptation)
- Fine-tuning hyperparameters were taken from UAlpaca
- Build diverse experimental setups: lean, full and multiple

Evaluation metrics: Mean Reciprocal Rank, Mean Average Precision, P@k

+ Mean Overlap Coefficient



Ukrainian Hypernym Discovery dataset

compiled from Ukrainian WordNet Basis

- Concept — common nouns
- Entity — specific persons, countries, and geographic entities

Lean Approach

- Generate six hypernyms for the word "input_term"

1 hypernym pattern=2590 samples

Full Setup

- Which terms belong to a higher abstraction level than "input_term"?
- Are there other general categories to which "input_term" can be attributed?

19 hypernym pattern=47,310 samples

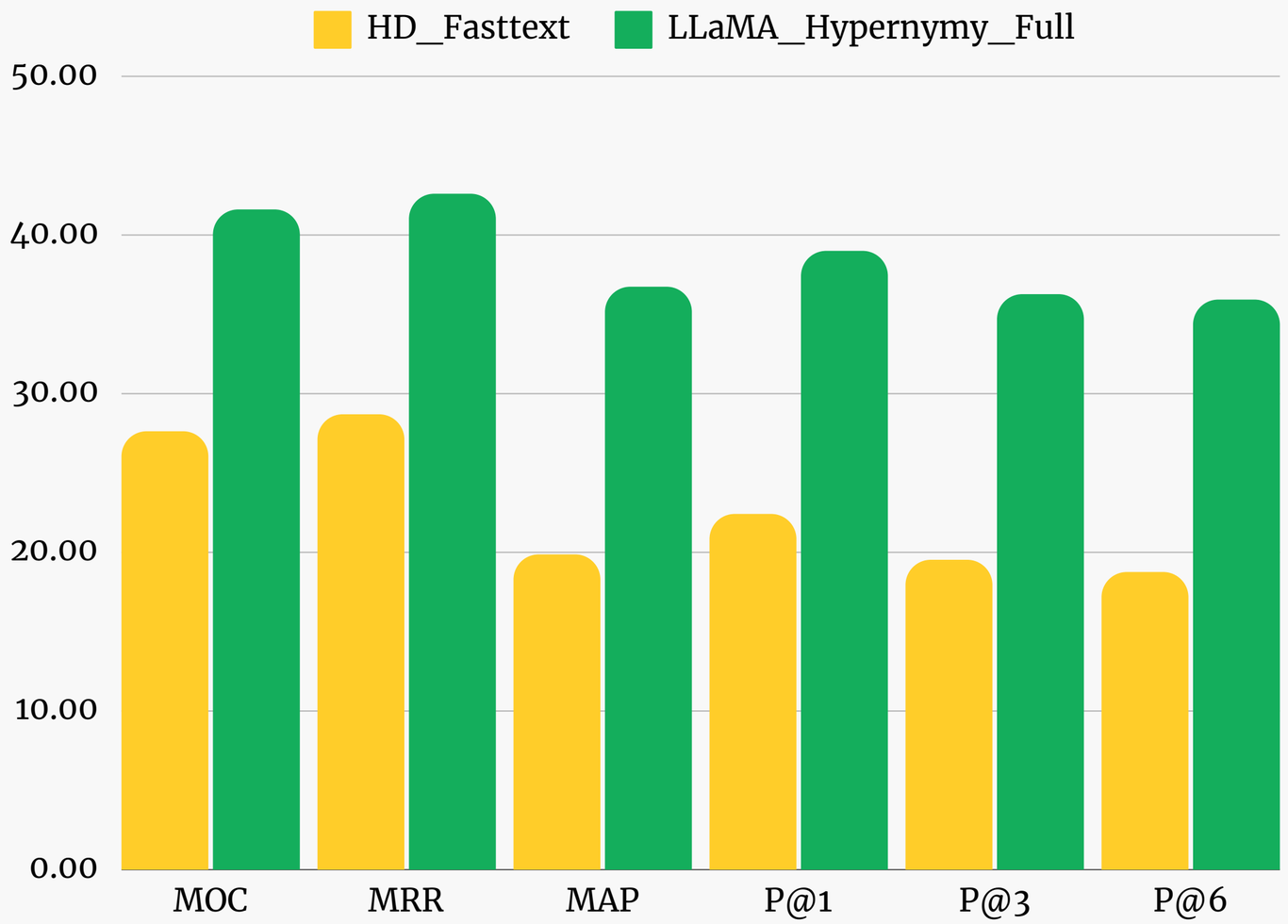
Multiple Relations

- What are other co-hyponyms for "input_term"?
- What concepts are more specific than "input_term?"

19 hypernym + 13 co-hyponyms + 14 hyponym pattern = 78,149 samples

Metrics comparison

of the two top-performing models



Error analysis

Examples of predictions made by the HD_Fasttext and LLaMA_Hypernymy_Full models

Query	OC	HD_Fasttext Predictions
		High OC examples
молюски shellfish	67	організм, їжа, безхребетні, дистрикт, артефакт, рослини organism, food, invertebrates, district, artifact, plants
Ориноко Orinoco	100	організм, дистрикт, муніципалітет, метрополіс, артефакт, річка organism, district, municipality, metropolis, artifact, river
		Low OC examples
Гіпатія Hypatia	17	дистрикт, організм, муніципалітет, артефакт, їжа, метрополіс district, organism, municipality, artifact, food, metropolis
Сапфо Sappho	0	метрополіс, артефакт, організм, дистрикт, муніципалітет, їжа metropolis, artifact, organism, district, municipality, food
Query	**OC**	**LLaMA_Hypernymy_Full Predictions**
		High OC examples
холангіт cholangitis	100	симптом, запалення, хвороба symptom, inflammation, disease
Неккар Neckar	100	річка river
		Low OC examples
метамфетамін methamphetamine	0	опіати, наркотик, анальгетики opiates, narcotic, analgesics
Сент-Джонс St. John's	0	озеро, річка lake, river

V Contribution

- Proposed a data-driven approach for automated hypernym hierarchy construction for the Ukrainian WordNet
- Suggested different techniques to generate candidates to fill the gaps
- Adapted SemEval 2018 Task 9: Hypernym Discovery to the Ukrainian language
- Explored the capabilities of SOTA LLMs for solving the Hypernym Discovery task
- Established a scalable foundation for creating a comprehensive and reliable WordNet for the Ukrainian language

---

All artifacts of this work, including code and data, are available on GitHub and HuggingFace.

V Future Work

I

Manually verify the automatically constructed concepts and relations

II

Extend the proposed solution to phrases, other semantic relations, and other parts of speech

III

Rerun the linking algorithm of Wikidata and Ukrainian Wikipedia to get more initial pairs

IV

Create a high-quality and comprehensive manual for annotators

V

WordNet should have a user-friendly interface accessible to the general public and linked to the OMW (Open Multilingual WordNet)

Nataliia Romanyshyn
Ukrainian Catholic University
romanyshyn.n@ucu.edu.ua

DMYTRO CHAPLYNSKYI
lang-uk
chaplinsky.dmitry@gmail.com

MARIANA ROMANYSHYN
Grammarly
mariana.romanyshyn@grammarly.com

Thank you for listening!

Time for Q&A