Taxonomy Enrichment without candidates

NLP
DL

Final Project

Skoltech

the team

Alsu Vakhitova

Skoltech, DS
Alsu.Vakhitova@skoltech.ru

Andreea Dogaru

Skoltech, DS
Andreea.Dogaru@skoltech.ru

Timotei Ardelean

Skoltech, DS
Timotei.Ardelean@skoltech.ru

Gabriel Rozzonelli

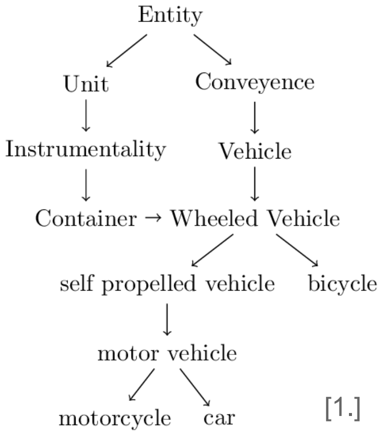
Skoltech, DS
Gabriel.Rozzonelli@skoltech.ru

Skoltech

Overview

- Motivation
- Problem Description
- Dataset
- Approaches
- Results
- Conclusion

Motivation



Entity
Unit
Conveyence
Instrumentality
Vehicle
Container → Wheeled Vehicle
self propelled vehicle
bicycle
motor vehicle
motorcycle
car

[1.]

- **Lexical resources**, such as WordNet [2.], are **important for the NLP community**
- Such datasets are **static**, when languages are naturally **dynamic**
- **Updating** them is rather **costly**

Can we use language models to find suitable candidates for enriching lexicons?

- Recent breakthroughs in the area of language models
- Rather efficient at addressing **masked token prediction**

```
[CLS] this project is [MASK] [SEP]
```

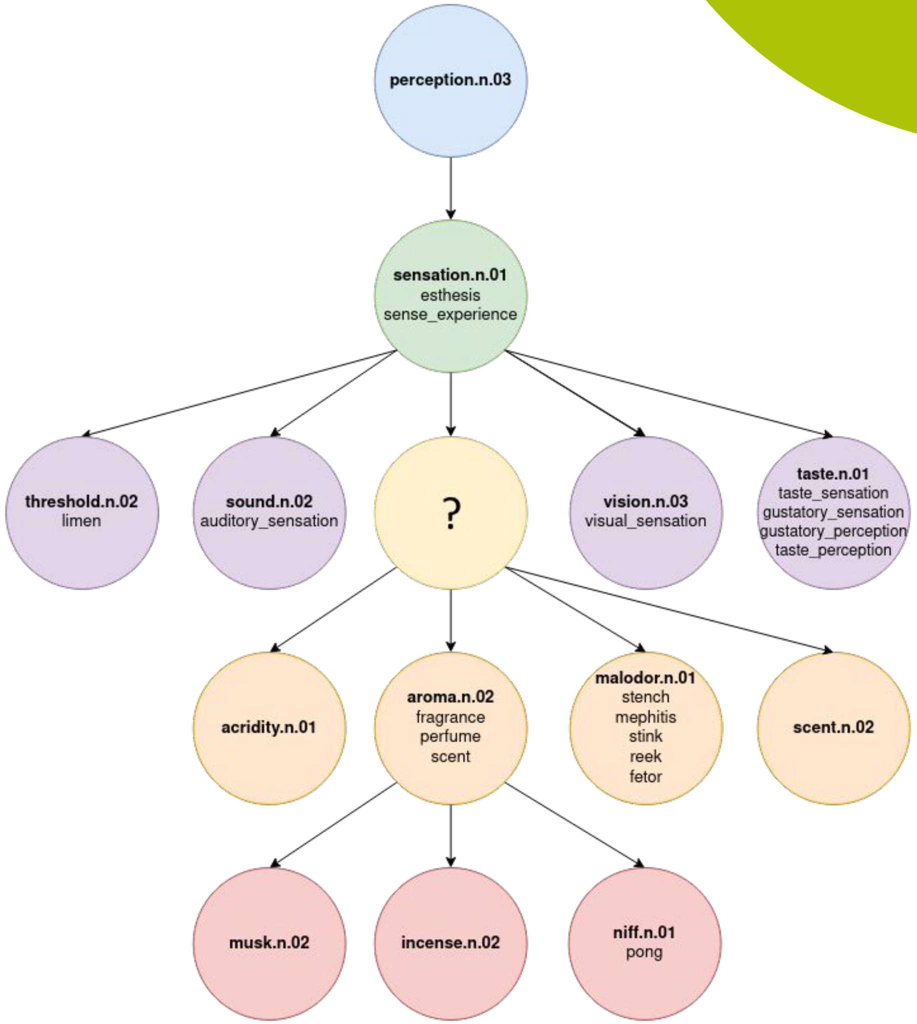
```
amazing
breathtaking
...
```

Problem description

- Taxonomies can be represented as **graphs**
  - Nodes → **synsets**
  - Edges → **hypernymies** (*is-a* relationships)
- Each synset has **lemmas**

Given that a synset lies at a certain position in the taxonomy, what are the lemmas which are more likely to be part of it?

WordNet subgraph

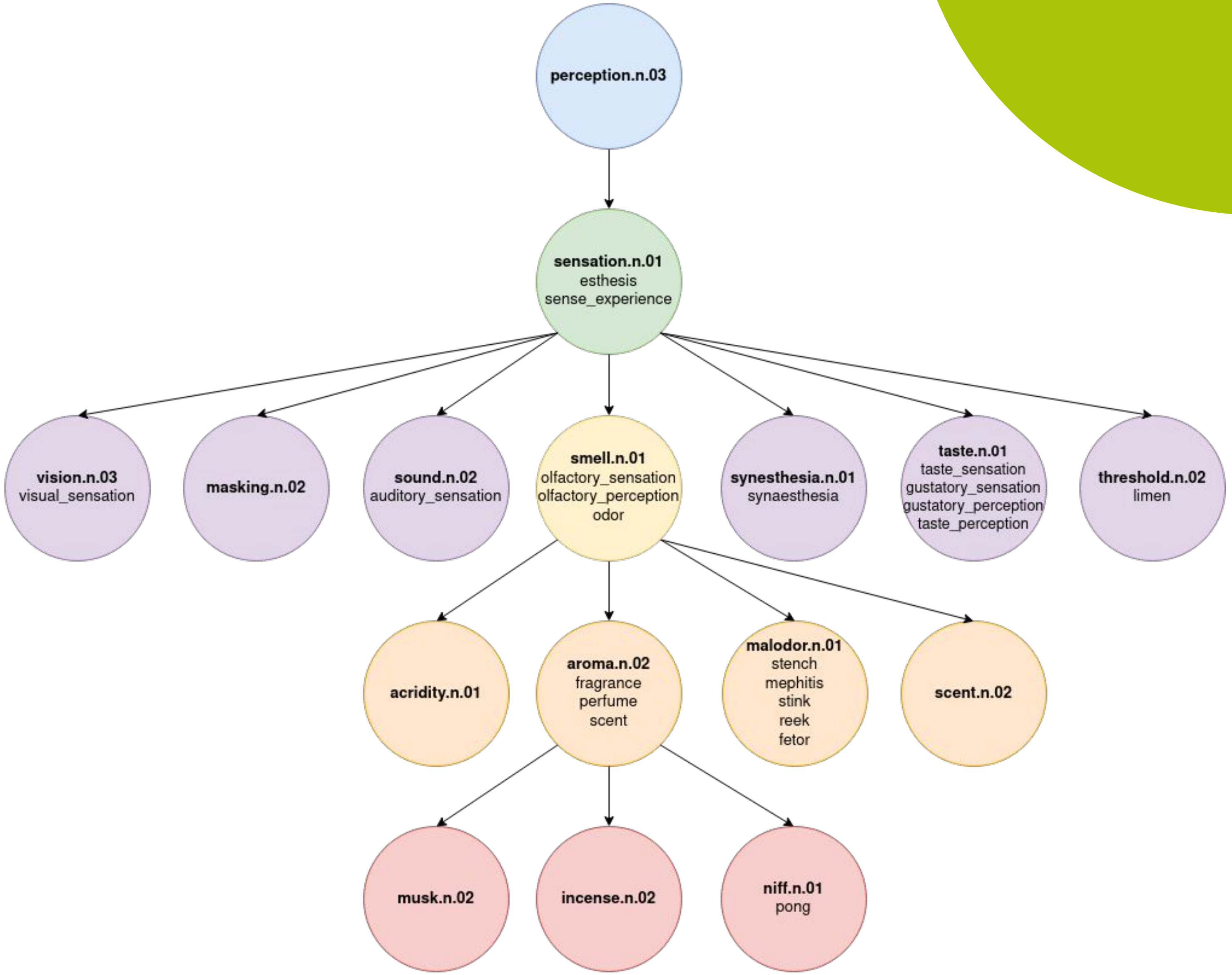


Skoltech

Dataset

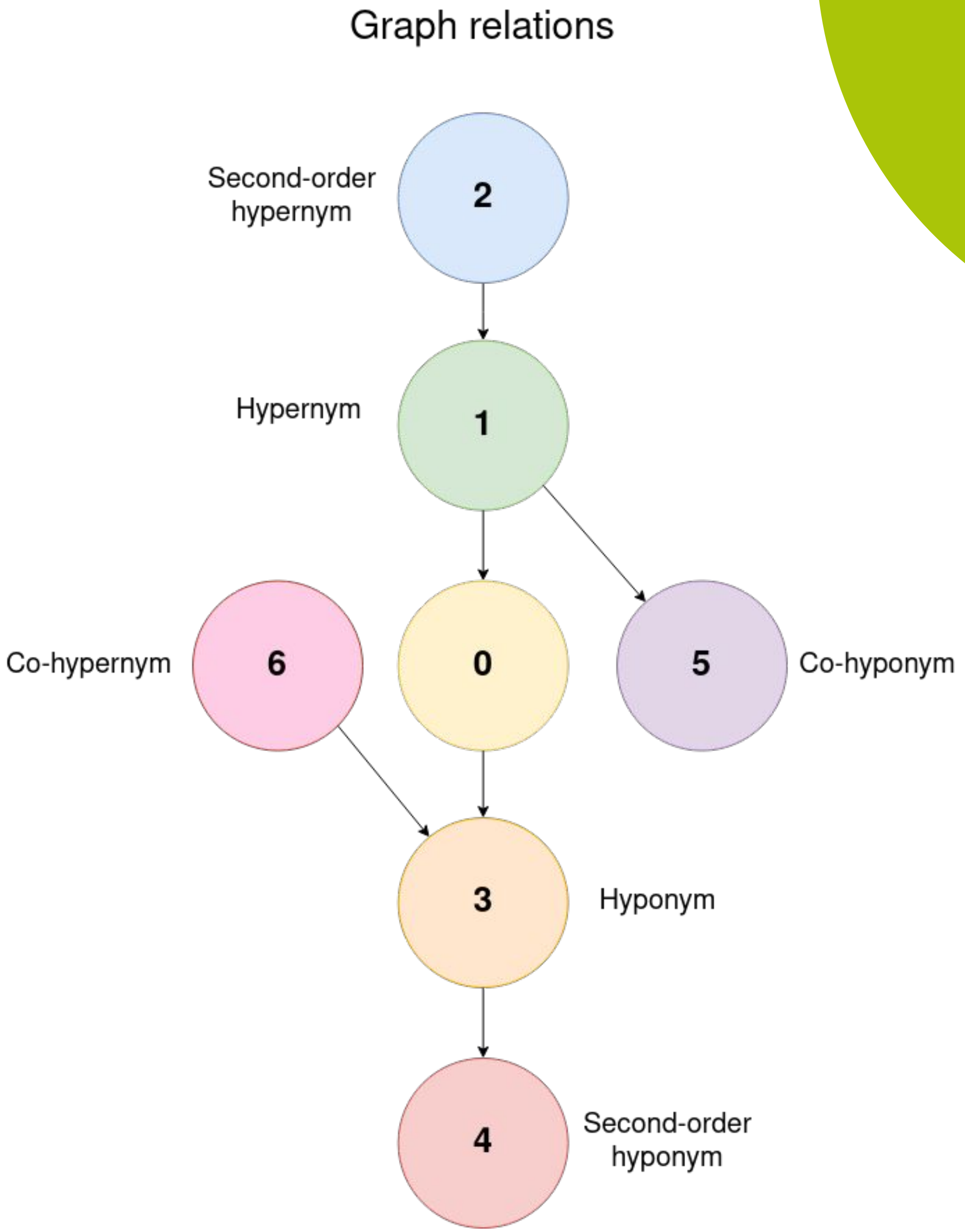
- Based on **WordNet** [2.] taxonomy
  - Train: 70999 entries
    - of which 10% for validation
  - Test: 3375 entries
- Subgraphs are centered on the target (query) node

WordNet subgraph



Dataset

An entry contains the lemmas for the target node, the lemmas of the neighboring synsets, and the graph relations between the nodes in the subgraph



Input data representation

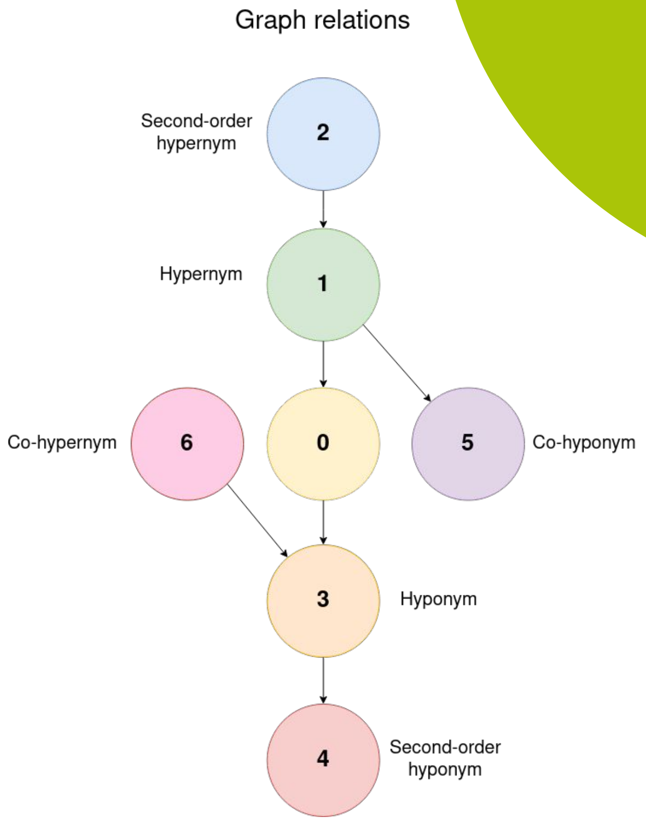
- **Token IDs**

  produced by the tokenizer; vocabulary indices of the word pieces
- **Level IDs**

  position relative to the central node within the taxonomic subgraph
- **Synset IDs**

  mark the appartenance of tokens to a particular synset
- **Highway**

  boolean indicator for tokens that belong to a synset name

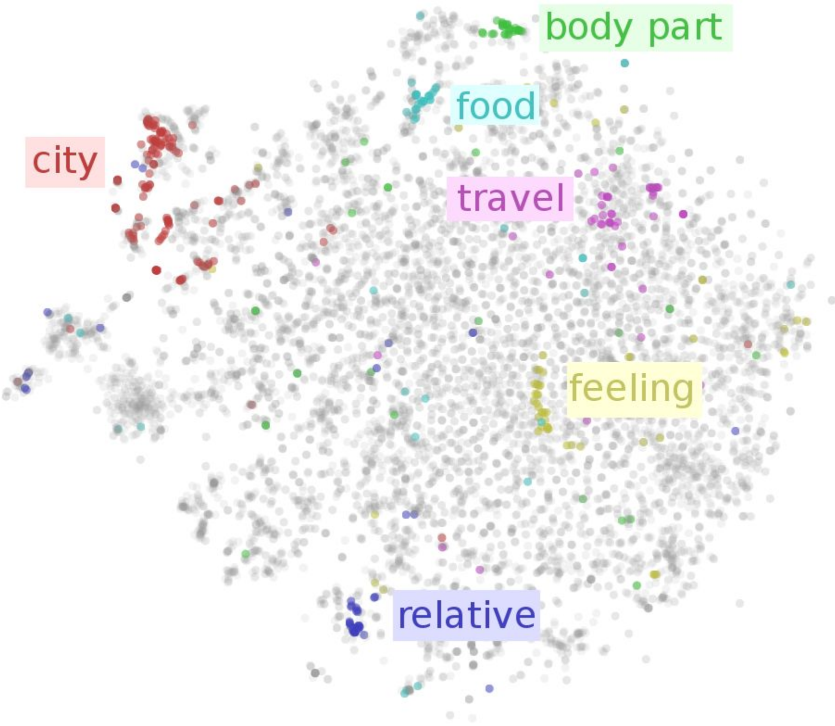


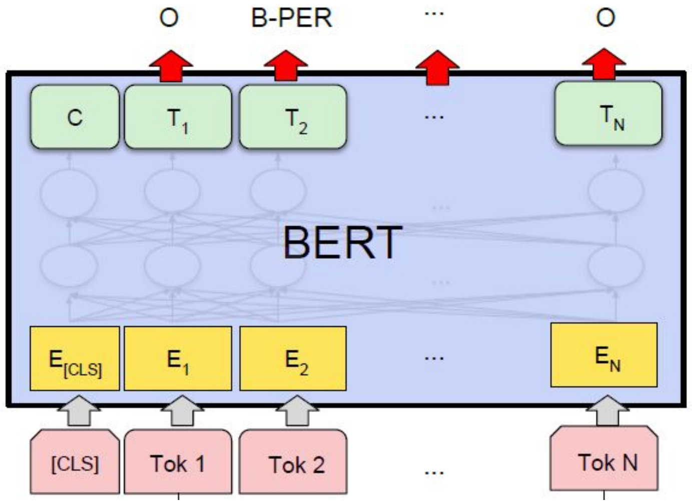
Tokens	[MASK]	[MASK]	[MASK]	sensation	est	##hesis	sense	experience	perception	aroma	fragrance	perfume	scent	mu	#sk	incense	ni	#ff	po	##ng	ac	#rid	##ity
Token IDs	103	103	103	8742	9765	24124	3168	3325	10617	23958	24980	17013	6518	14163	6711	28647	9152	4246	13433	3070	9353	14615	3012
Level IDs	0	0	0	1	1	1	1	1	2	3	3	3	3	4	4	4	4	4	4	4	3	3	3
Synset IDs	0	0	0	1	1	1	1	1	2	3	3	3	3	4	4	5	6	6	6	6	7	7	7
Highway	1	1	1	1	0	0	0	0	1	1	0	0	0	1	1	1	1	1	0	0	1	1	1

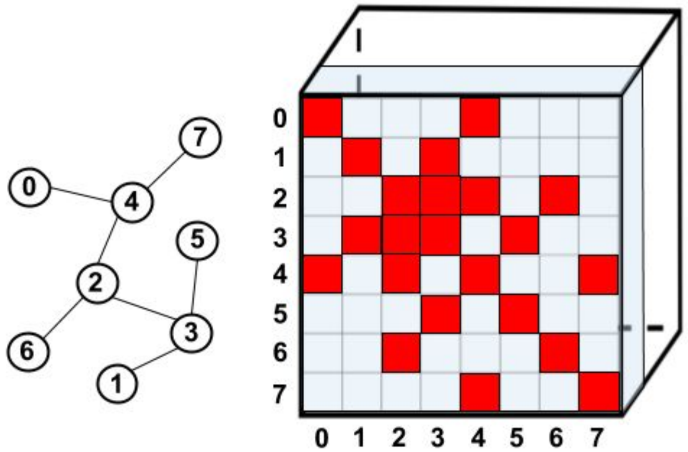
...

Approaches

- Fixed-Vocabulary Baseline
- KBERT
- KBERT + GAT





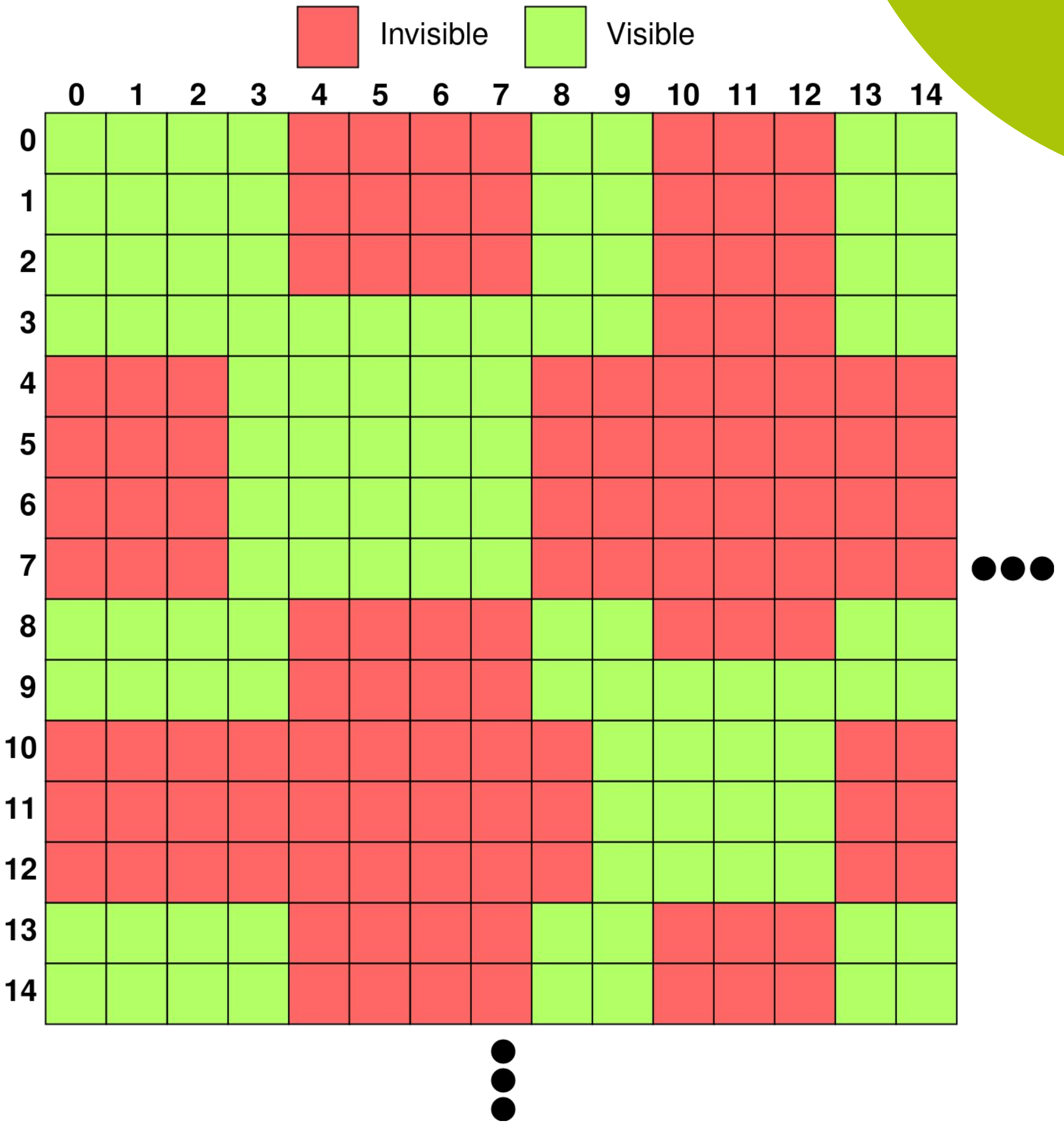


Baseline

- Uses a fixed vocabulary for possible lemmas suggestions
- Relies on word embeddings pre-trained on large corpora to represent the meaning.
- Tasked to predict the embedding for the query node in the taxonomy
- Ranks the words in the taxonomy based on cosine similarity

KBERT

- Based on BERT masked language model
- Supports enriched input data with additional lemmas of the neighboring nodes
- Prevents knowledge noise issue through a “visibility matrix” that restricts attention in the Multi-Head Attention layers



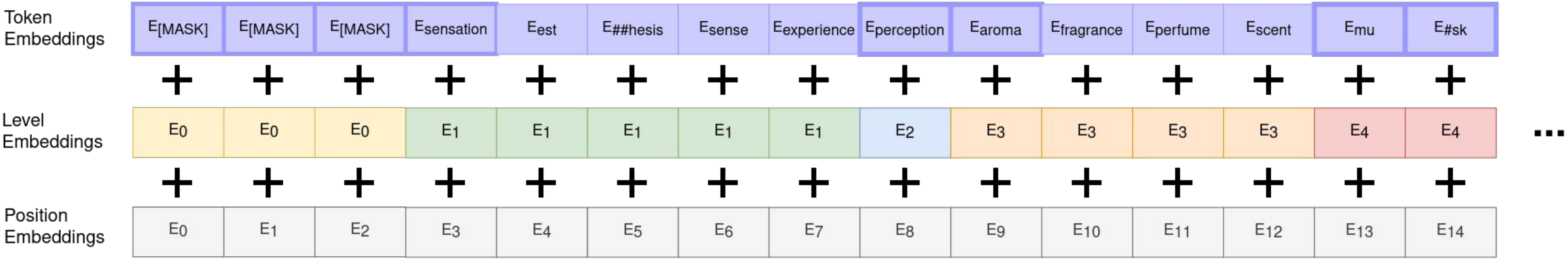
🟥 Invisible 🟩 Visible

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	🟩	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟩	🟩	🟥	🟥	🟥	🟩	🟩
1	🟩	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟩	🟩	🟥	🟥	🟥	🟩	🟩
2	🟩	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟩	🟩	🟥	🟥	🟥	🟩	🟩
3	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟥	🟥	🟩	🟩
4	🟥	🟥	🟥	🟩	🟩	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥
5	🟥	🟥	🟥	🟩	🟩	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥
6	🟥	🟥	🟥	🟩	🟩	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥
7	🟥	🟥	🟥	🟩	🟩	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥
8	🟩	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟩	🟩	🟥	🟥	🟥	🟩	🟩
9	🟩	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩
10	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟩	🟩	🟩	🟥	🟥
11	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟩	🟩	🟩	🟥	🟥
12	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟩	🟩	🟩	🟥	🟥
13	🟩	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟩	🟩	🟥	🟥	🟥	🟩	🟩
14	🟩	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟩	🟩	🟥	🟥	🟥	🟩	🟩

●●●

KBERT

- TaxoEmbedder

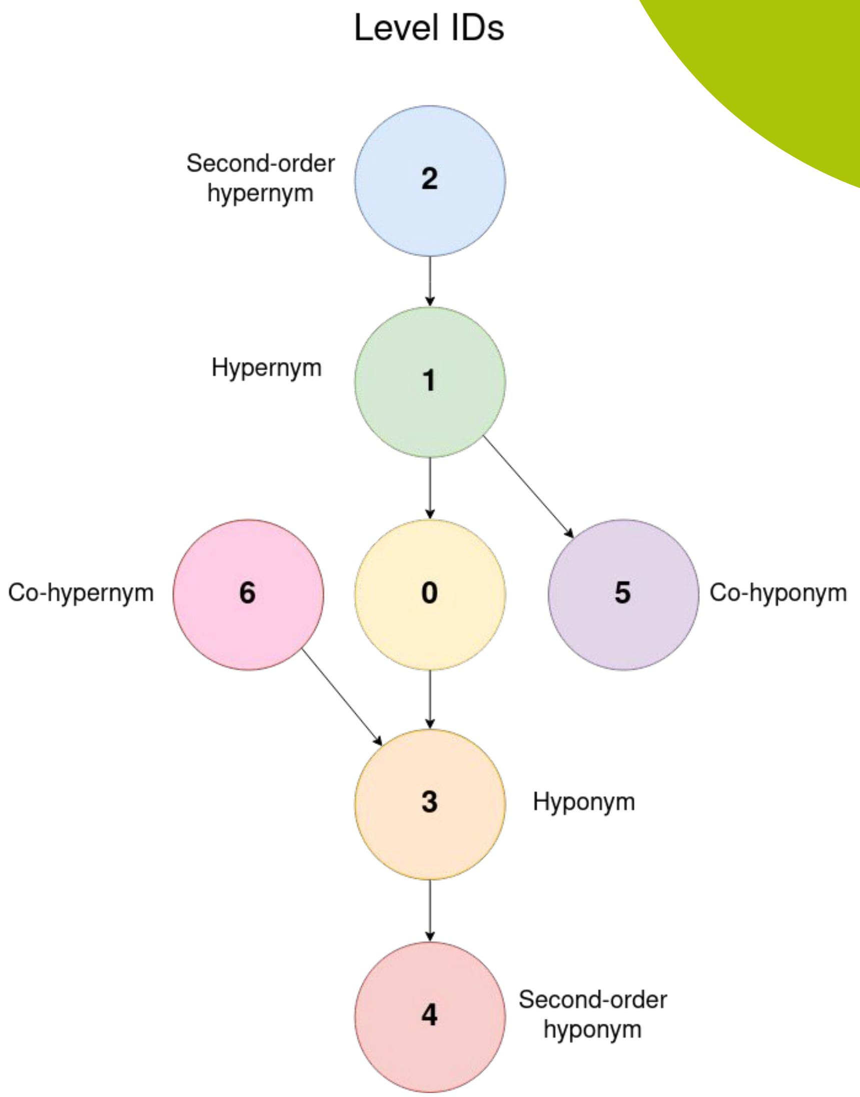


Token Embeddings	$E_{\text{[MASK]}}$	$E_{\text{[MASK]}}$	$E_{\text{[MASK]}}$	$E_{\text{sensation}}$	E_{est}	$E_{\text{\#\#hesis}}$	E_{sense}	$E_{\text{experience}}$	$E_{\text{perception}}$	E_{aroma}	$E_{\text{fragrance}}$	E_{perfume}	E_{scent}	E_{mu}	$E_{\text{\#sk}}$	
	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
Level Embeddings	E_0	E_0	E_0	E_1	E_1	E_1	E_1	E_1	E_2	E_3	E_3	E_3	E_3	E_4	E_4	...
	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
Position Embeddings	E_0	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}	E_{11}	E_{12}	E_{13}	E_{14}	

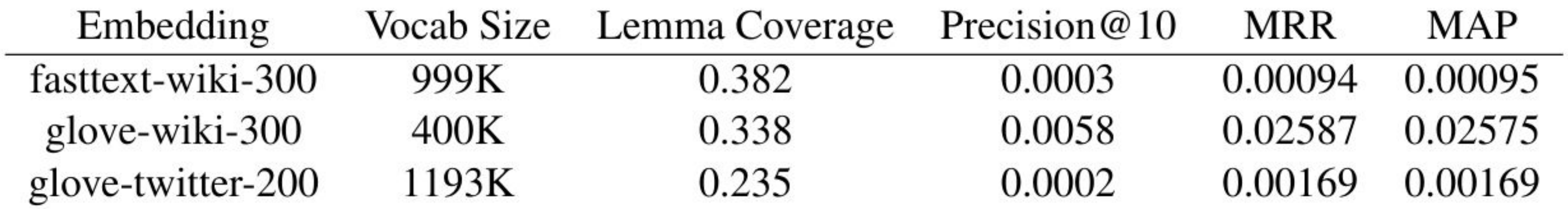
- BERT$_{\text{BASE}}$ encoder (12 layers with 768 hidden size)
- Classification head (2 linear transformations)

KBERT-GAT

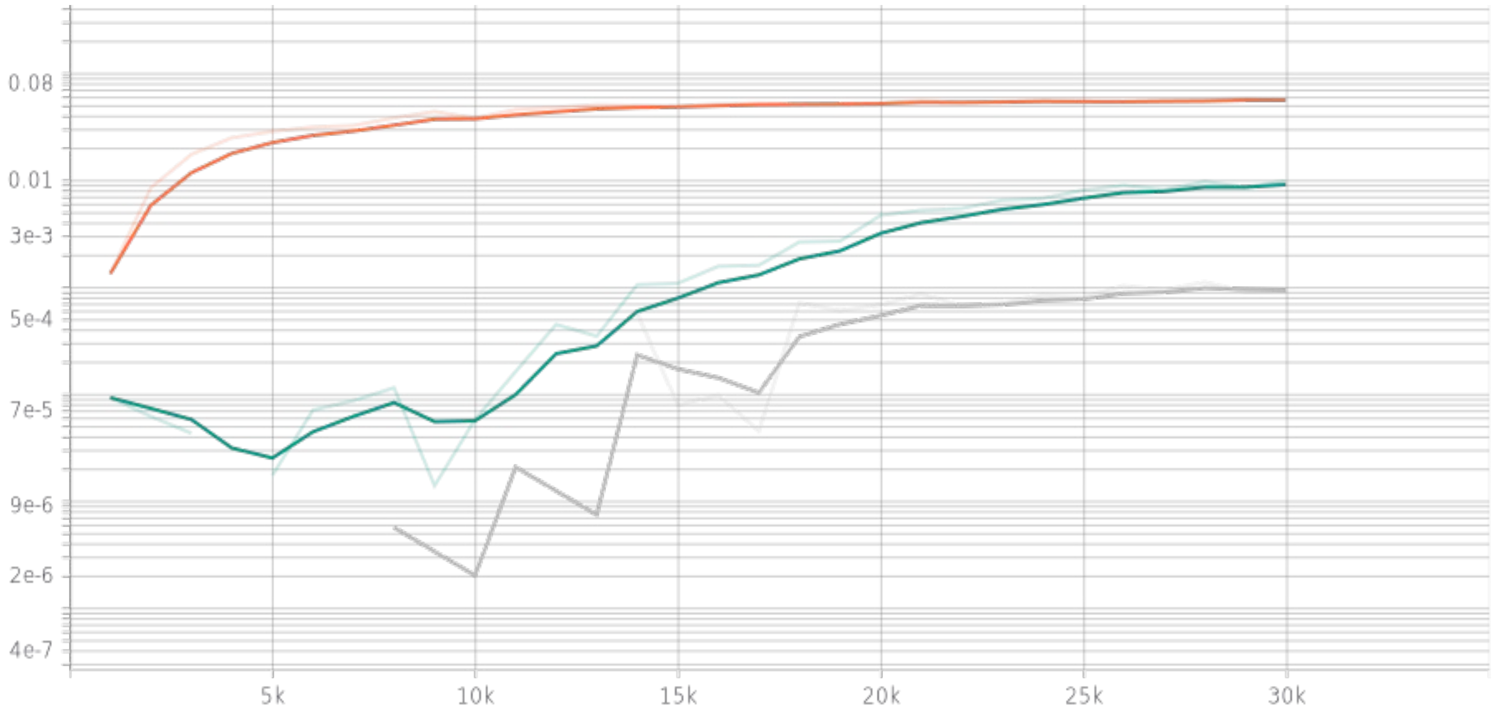
- Extends the K-BERT solution
  - Same embedder
  - Same encoder
- Replaced classification head with **Graph Attention Network (GAT)**
- Our novelty: use **graph visible matrix** instead of a simple adjacency matrix in a multi-head attention
  - All lemmas within *one synset* can attend each other
  - Only highway lemmas that have *adjacent levels* can attend each other



Baseline Results



Embedding	Vocab Size	Lemma Coverage	Precision@10	MRR	MAP
fasttext-wiki-300	999K	0.382	0.0003	0.00094	0.00095
glove-wiki-300	400K	0.338	0.0058	0.02587	0.02575
glove-twitter-200	1193K	0.235	0.0002	0.00169	0.00169



MRR (logscale) for different embeddings.

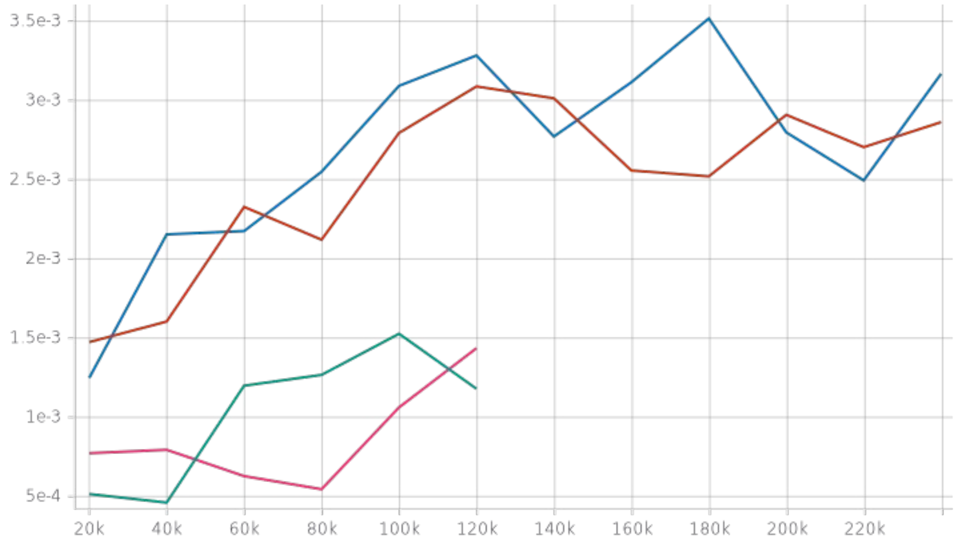
Legend:
Glove-Wiki-300,
Glove-Twitter-200,
Fasttext-300

KBERT Results

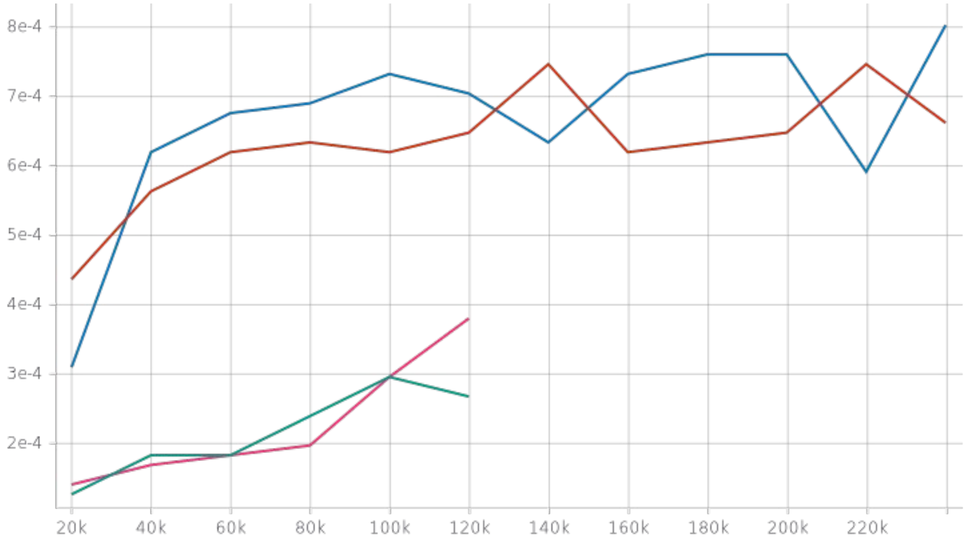
	Encoder	Head	Precision@10	MRR	MAP
🟩	-	-	0.00028	0.0010	0.0010
🟪	+	-	0.00030	0.0011	0.0011
🟥	+	+	0.00038	0.0011	0.0011
🟦	+	+*	0.00038	0.0014	0.0014

Legend:
− trained from scratch
+ pre-trained
✶ frozen

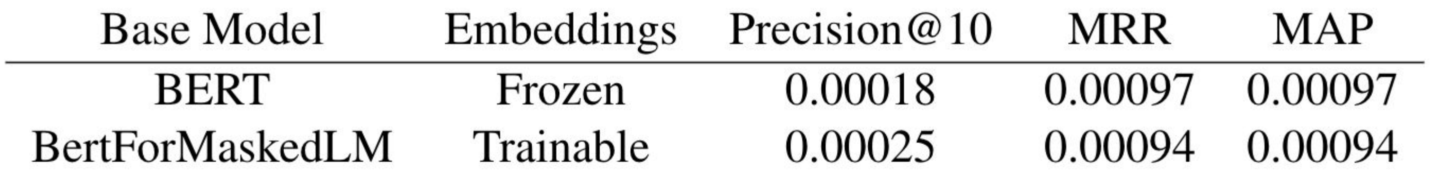
Validation MRR



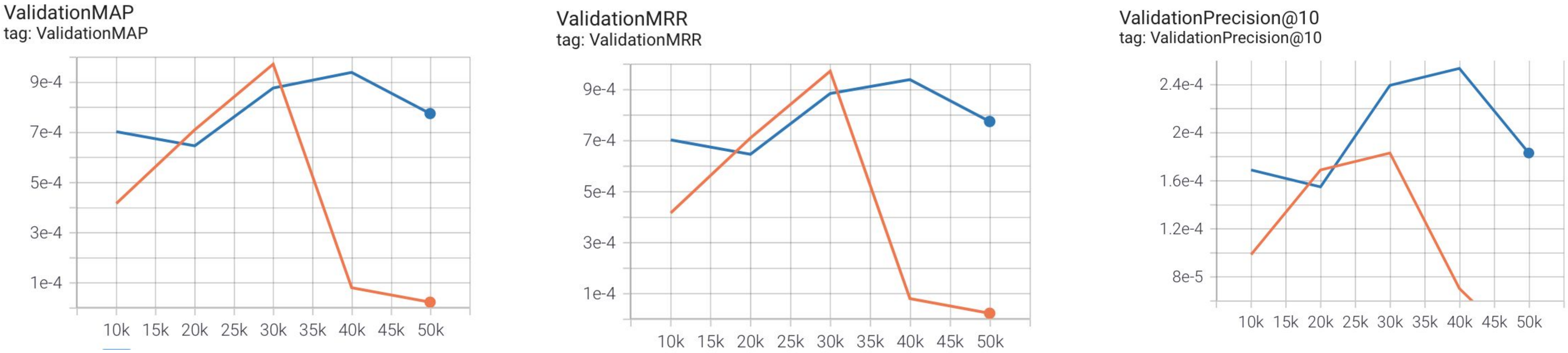
Validation Precision@10



KBERT-GAT Results



Base Model	Embeddings	Precision@10	MRR	MAP
BERT	Frozen	0.00018	0.00097	0.00097
BertForMaskedLM	Trainable	0.00025	0.00094	0.00094



Legend: **BERT**, **BertForMaskedLM**

Conclusion

- We propose a new approach to address the task of **taxonomy enrichment without candidates**
- In this regard, we implemented **two systems** based on KBERT: with and without GAT
- Results indicate that candidate-free taxonomy enrichment is **relevant** and **feasible**

thx.

Questions?

Skoltech

References

1. Atish Pawar, Vijay Mago. 2018. Calculating the similarity between words and sentences using a lexical database and corpus statistics.
2. George A Miller. 1998. WordNet: An electronic lexical database. MIT press.

Resources

- GitHub repository: https://github.com/palette-knife25/candidate-free-te