UNIVERSITÀ DI PISA

Regularizing Transformers By Symbolic Knowledge and Deep Graph Networks

Master's Degree in Computer Science

Author

Matteo Medioli

Supervisor

Prof. Dr. Davide Bacciu

Co-Supervisor

Andrea Valenti

Dr. Lucia Passaro

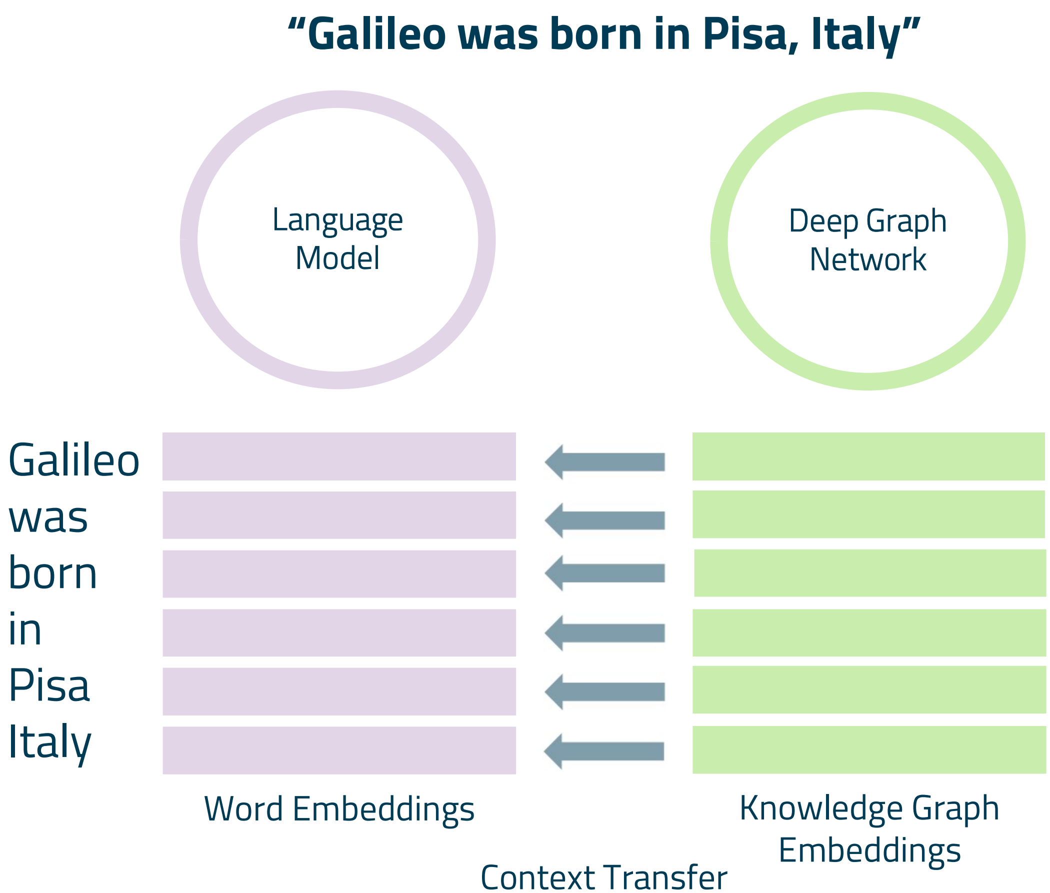
OUTLINE

- High Level Overview
- Language Models
- Knowledge Graphs
- KB Graph Attention Network
- Map Words to Node Embeddings
- Masked Language Modeling Regularization Term
- Experiments
- Results

HIGH LEVEL OVERVIEW

CORPUS

Galileo was born in Pisa, Italy, on 15 February 1564, the first of six children of Vincenzo Galilei, a lutenist, composer, and music theorist, and Giulia Ammannati, who had married in 1562. Galileo studied speed and velocity, gravity and free fall, the principle of relativity, inertia, projectile motion and also worked in applied science and technology, describing the properties of pendulums and "hydrostatic balances". He invented the thermoscope and various military compasses, and used the telescope for scientific observations of celestial objects. His contributions to observational astronomy include telescopic confirmation of the phases of Venus, observation of the four largest satellites of Jupiter, observation of Saturn's rings, and analysis of lunar craters and sunspots.



HIGH LEVEL OVERVIEW

CHALLENGES

Define two representations of words in text corpus: **word embeddings** and **node embeddings**

Define **regularization term** $\mathcal{L}_R$ in MLM task to transfer symbolic knowledge of **knowledge graphs** to **language models**

Train from scratch **BERT-KG** and **BERT-BASELINE** to analyze the influence of regularization through **Word Embeddings Probing Tasks**

HIGH LEVEL OVERVIEW

THESIS CONTRIBUTION

DGN in Natural Language Processing

- VGAE (*Kipf and Welling, 2016*)
- Text-GCN (*Yao et al, 2018*)
- VGCN-BERT (*Lu et al, 2020*)

Adapts Different Embedding Spaces

- ALC Embedding (*Khodak et al, 2018*)
- DMN Embedding (*Ni et al, 2018*)

Fine Tuning Approaches

⇓

Proposed Framework
Masked Language Modeling from Scratch

Context Transfer on Same Embedding Types

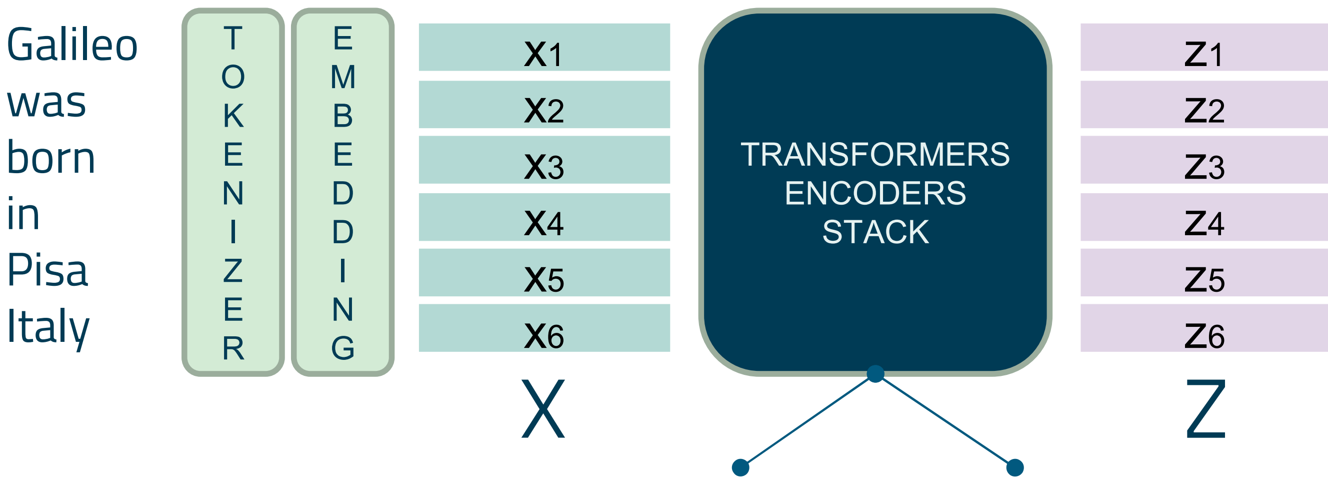
⇓

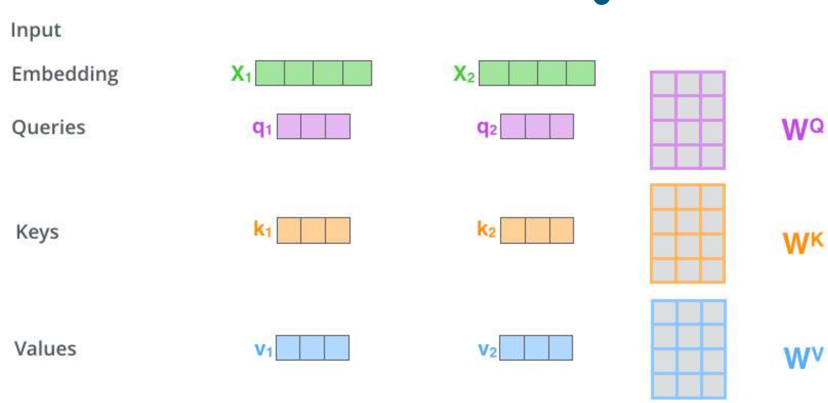
Proposed Framework
Context Transfer from
Node Embeddings to Word Embeddings

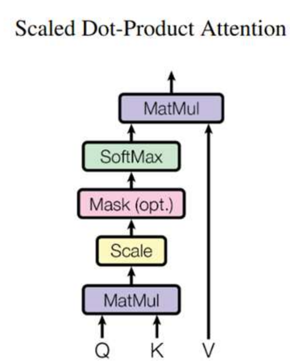
LANGUAGE MODELS

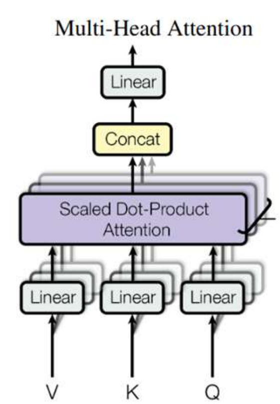
BERT

Bidirectional Encoder Representations from Transformers (*Devlin et al, 2018*)



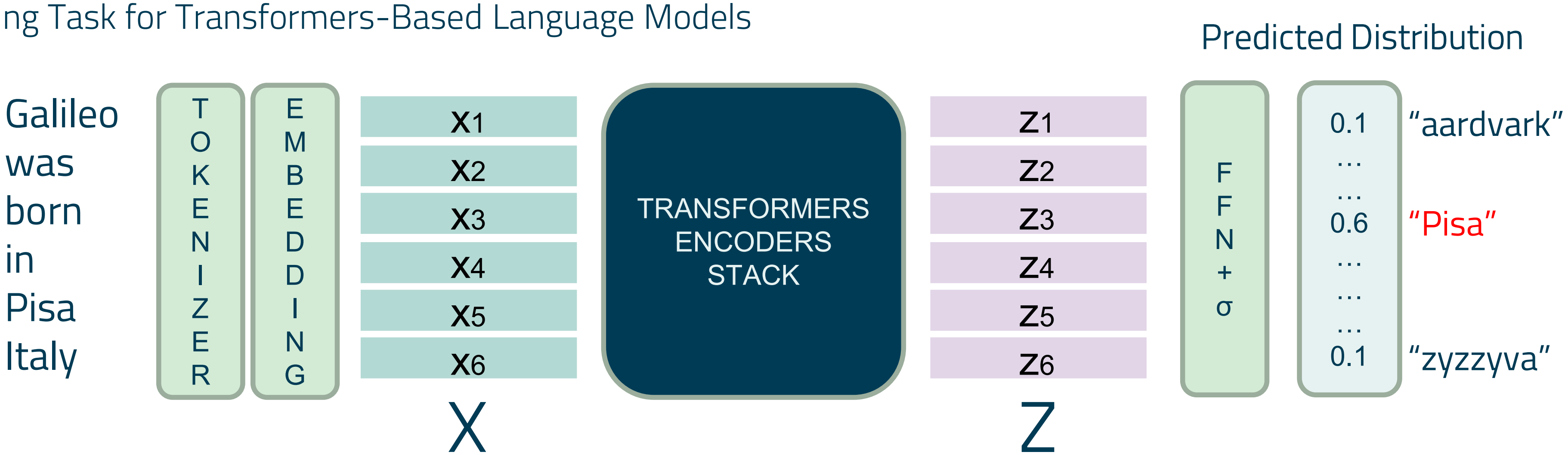






MASKED LANGUAGE MODELING

Training Task for Transformers-Based Language Models



Training Objective

$$\mathcal{L}_{\mathrm{MLM}}\left(\mathbf{X}_{\Pi} \mid \mathbf{X}_{-\Pi}, \theta\right) = \frac{1}{K} \sum_{k=1}^{K} \log p\left(\mathbf{x}_{\pi_k} \mid \mathbf{X}_{-\Pi}; \theta\right)$$

KNOWLEDGE GRAPHS

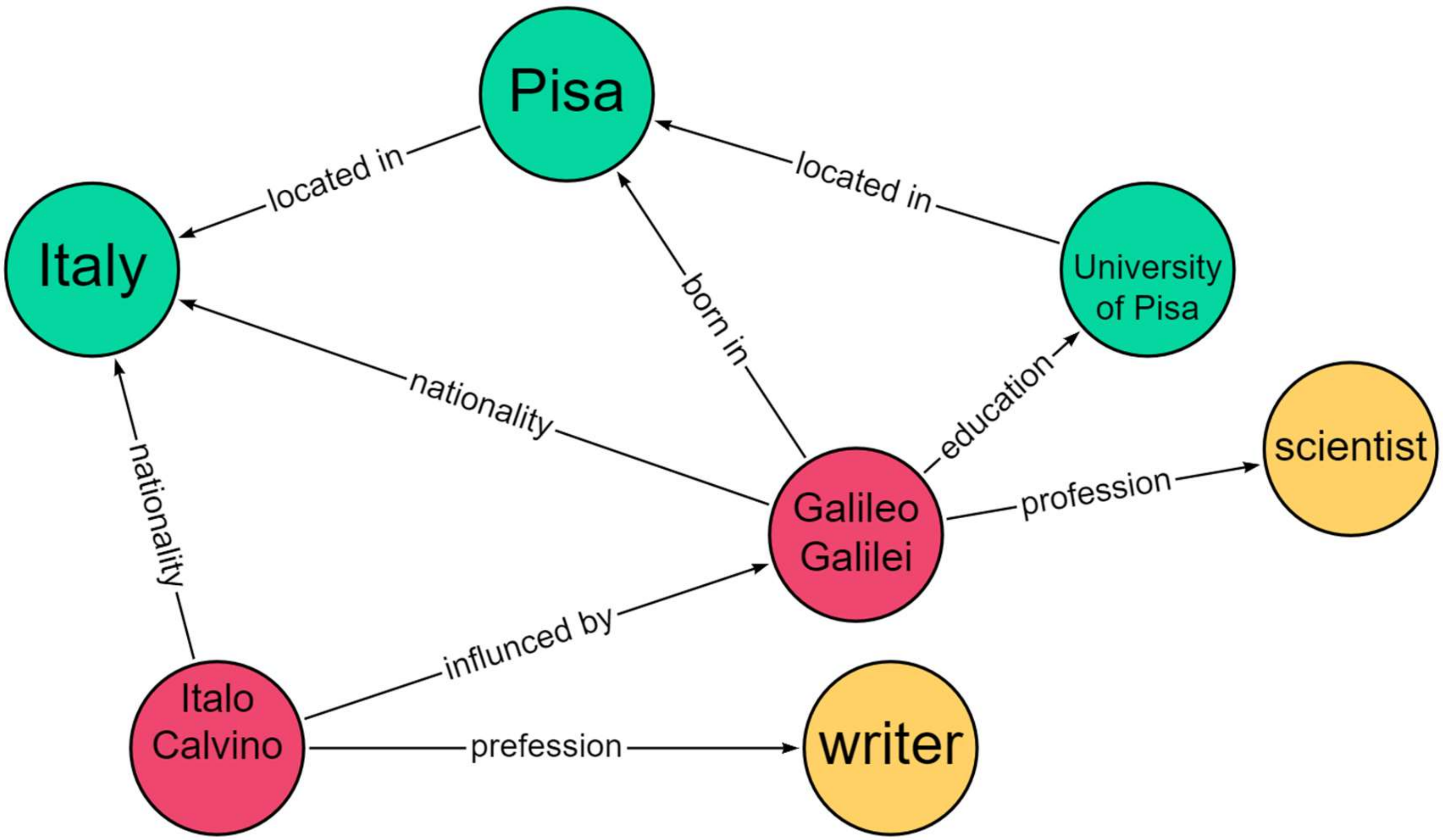
RDF GRAPHS

RDF TRIPLE

- Subject e_i
- Predicate r_k
- Object e_j

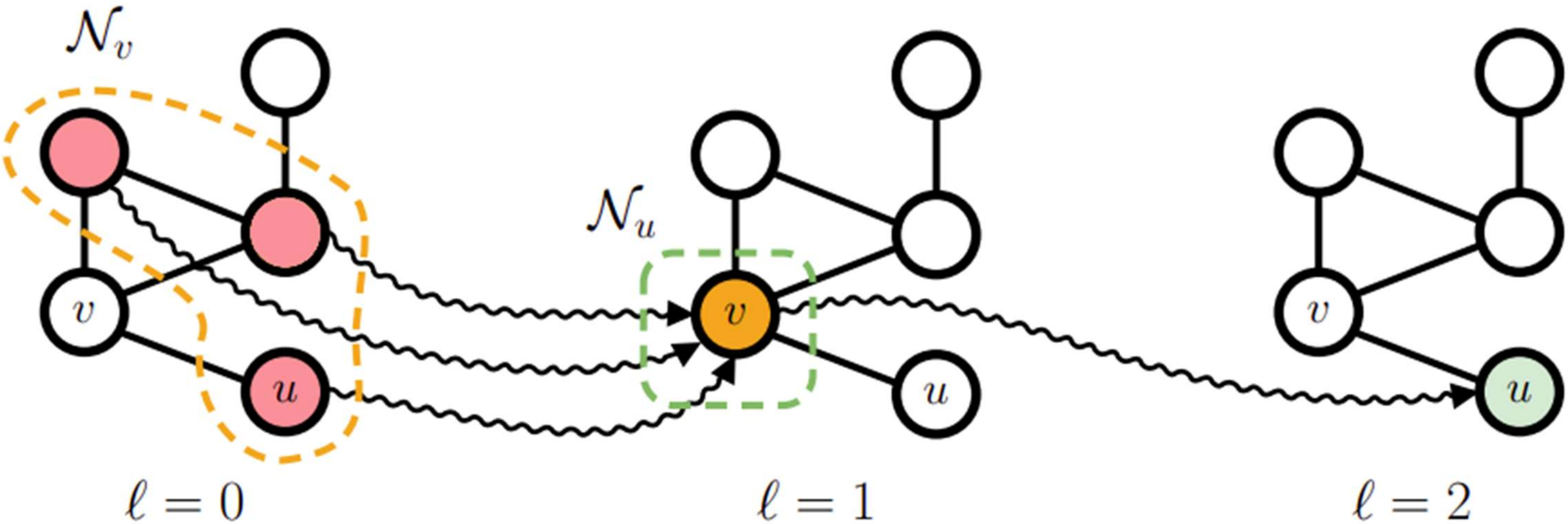
- Entities: e_i, e_j
- Relationships: r_k
- Facts and Concepts: (e_i, r_k, e_j)
- Ontology with Domain D

KNOWLEDGE GRAPH

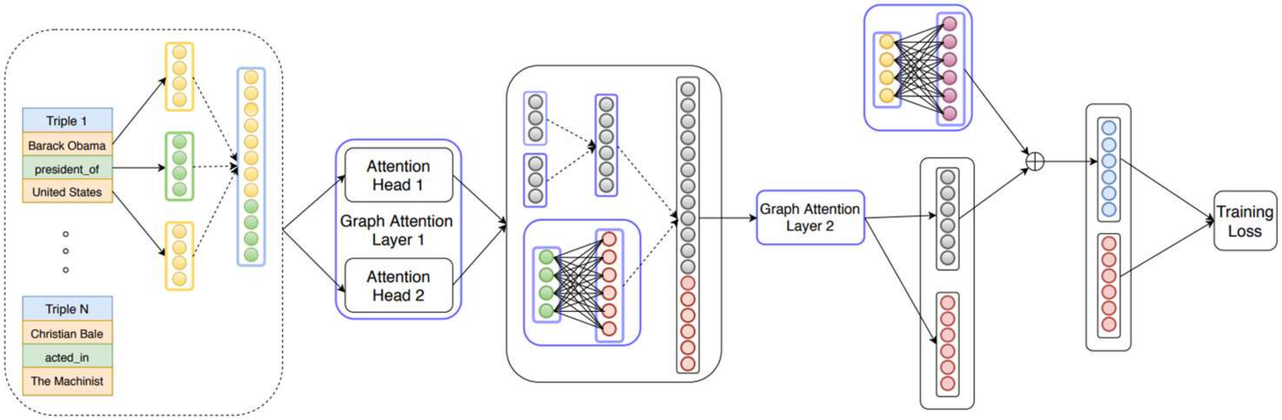


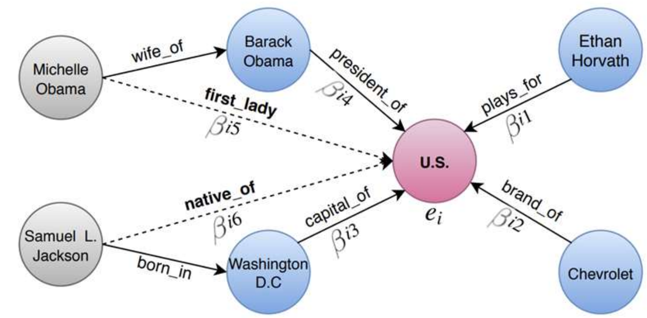
DEEP GRAPH NETWORKS

MESSAGE PASSING



KB GRAPH ATTENTION NETWORK

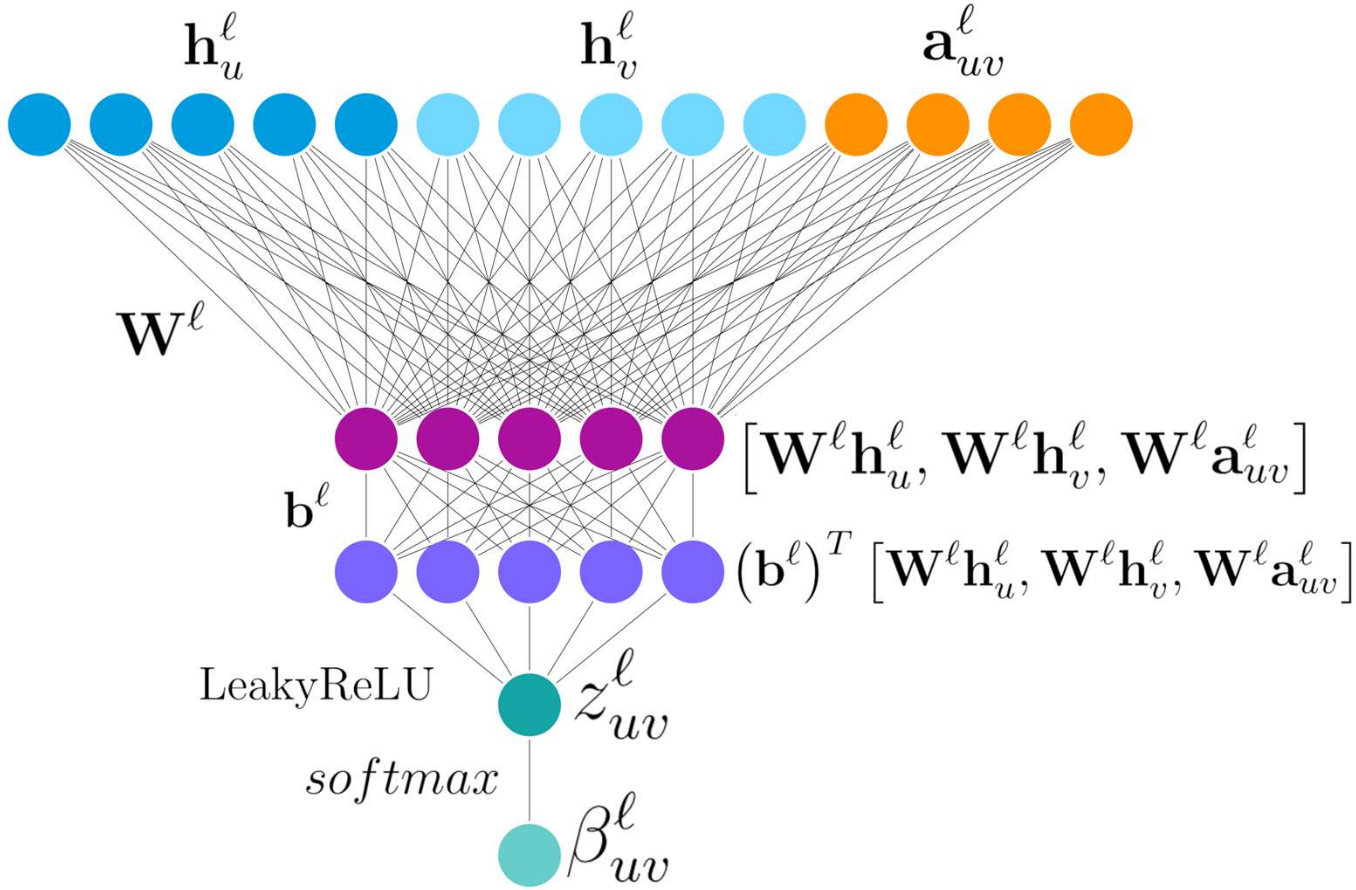




KBGAT Neighborhood Aggregation Rule

$$\mathbf{h}_v^{\ell+1} = \sigma\left(\sum_{u\in\mathcal{N}_v} \beta_{uv}^{\ell+1} * \left[\mathbf{W}^{\ell+1}\mathbf{h}_u^{\ell}, \mathbf{W}^{\ell+1}\mathbf{h}_v^{\ell}, \mathbf{W}^{\ell+1}\mathbf{a}_{uv}^{\ell}\right]\right)$$

GRAPH ATTENTION LAYER



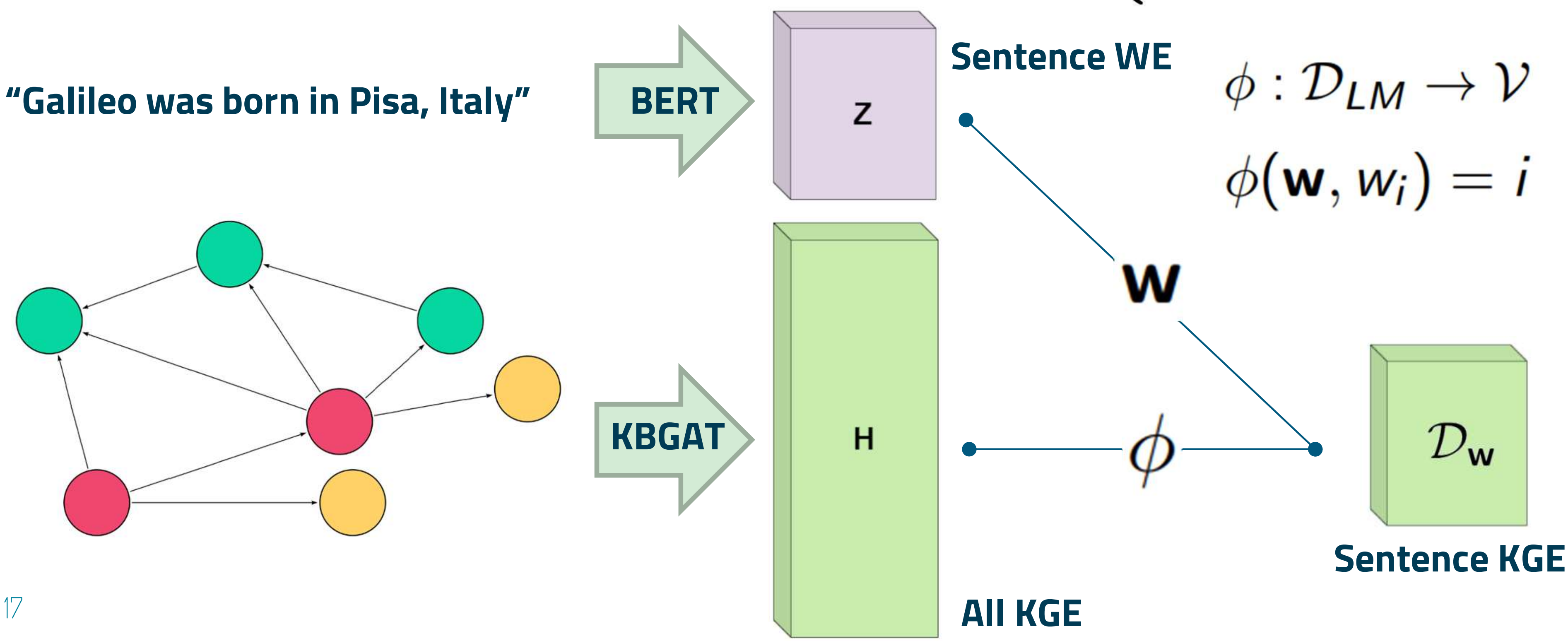
MAP WORDS TO NODE EMBEDDINGS

KGE SENTENCE DICTIONARY

$$\mathcal{D}_{\mathbf{w}}[k \rightarrow v](w_i) = \begin{cases} \mathbf{h}_{\phi(\mathbf{w}, w_i)} & \text{for } w_i = k \\ \mathbf{h}_{none} & \text{otherwise} \end{cases}$$

$$\phi : \mathcal{D}_{LM} \rightarrow \mathcal{V}$$

$$\phi(\mathbf{w}, w_i) = i$$



WORD SENSE DISAMBIGUATION

$\phi : \mathcal{D}_{LM} \rightarrow \mathcal{V}$ **as WSD Problem**

"The bank will not be accepting cash on Saturday"

"The river overflowed the bank"

Maximize the similarity function ψ between the candidate nodes $\mathcal{S}_i$ for the target word w_i and candidates nodes $\mathcal{S}_j$ for all the other nodes in the sentence:

$$\phi(\mathbf{w}, w_i) = \max_{i \in \mathcal{S}_i} \sum_{w_j \in \mathbf{w}} \max_{j \in \mathcal{S}_j} \psi(i, j)$$

Wu-Palmer similarity $\delta_{wup}(i, j) = \frac{2d}{L_i + L_j + 2d}$

Similarity Function $\psi : \mathcal{V} \times \mathcal{V} \rightarrow \mathbb{R}$

Cosine similarity

MLM REGULARIZATION TERM

REGRESSION FOR CONTEXT TRANSFER

Given the input sentence $\mathbf{w} = \{w_i\}_{i=0}^{N}$

- $\mathbf{x}_i = Tokenizer(w_i)$
- $\mathbf{z}_i = Encoder(\mathbf{x}_i)$
- $\mathbf{h}_i = \mathcal{D}_{\mathbf{w}}(w_i)$
- $\mathbf{Z}_{\mathbf{w}} = \left[(\mathbf{z}_1)^T, \ldots, (\mathbf{z}_N)^T\right]^T$
- $\mathbf{H}_{\mathbf{w}} = \left[(\mathcal{D}_{\mathbf{w}}(w_1))^T, \ldots, (\mathcal{D}_{\mathbf{w}}(w_N))^T\right]^T = \left[(\mathbf{h}_1)^T, \ldots, (\mathbf{h}_N)^T\right]^T$

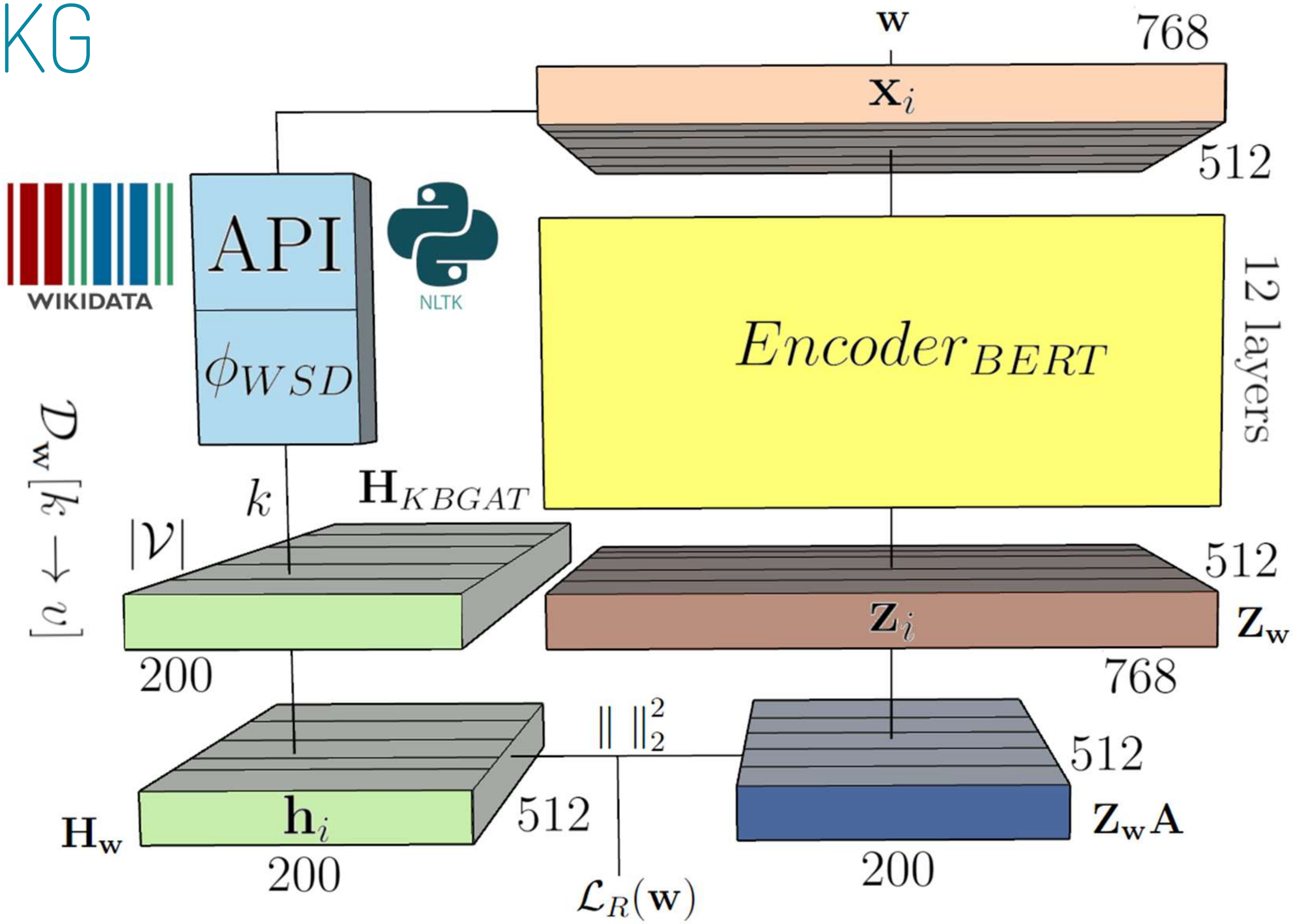
REGRESSION FOR CONTEXT TRANSFER

- Linear matrix $\mathbf{A} \in \mathbb{R}^{d_{hidden} \times d_{KGE}}$
- $\mathcal{L}_R(\mathbf{w}) = \sum_{w_i \in \mathbf{w}} \|\mathbf{h}_i - \mathbf{z}_i \mathbf{A}\|_2^2 = \|\mathbf{H}_\mathbf{w} - \mathbf{Z}_\mathbf{w}\mathbf{A}\|_2^2$

BERT−KG Training Loss:

$$\mathcal{L}_{\mathrm{MLM}}(\mathbf{X}_\Pi \mid \mathbf{X}_{-\Pi}, \theta) = \boxed{\frac{1}{K}\sum_{k=1}^{K} \log p(\mathbf{x}_{\pi_k} \mid \mathbf{X}_{-\Pi}; \theta)} + \boxed{\lambda \mathcal{L}_R(\mathbf{X}_{-\Pi}, \mathbf{H}, \theta)}$$

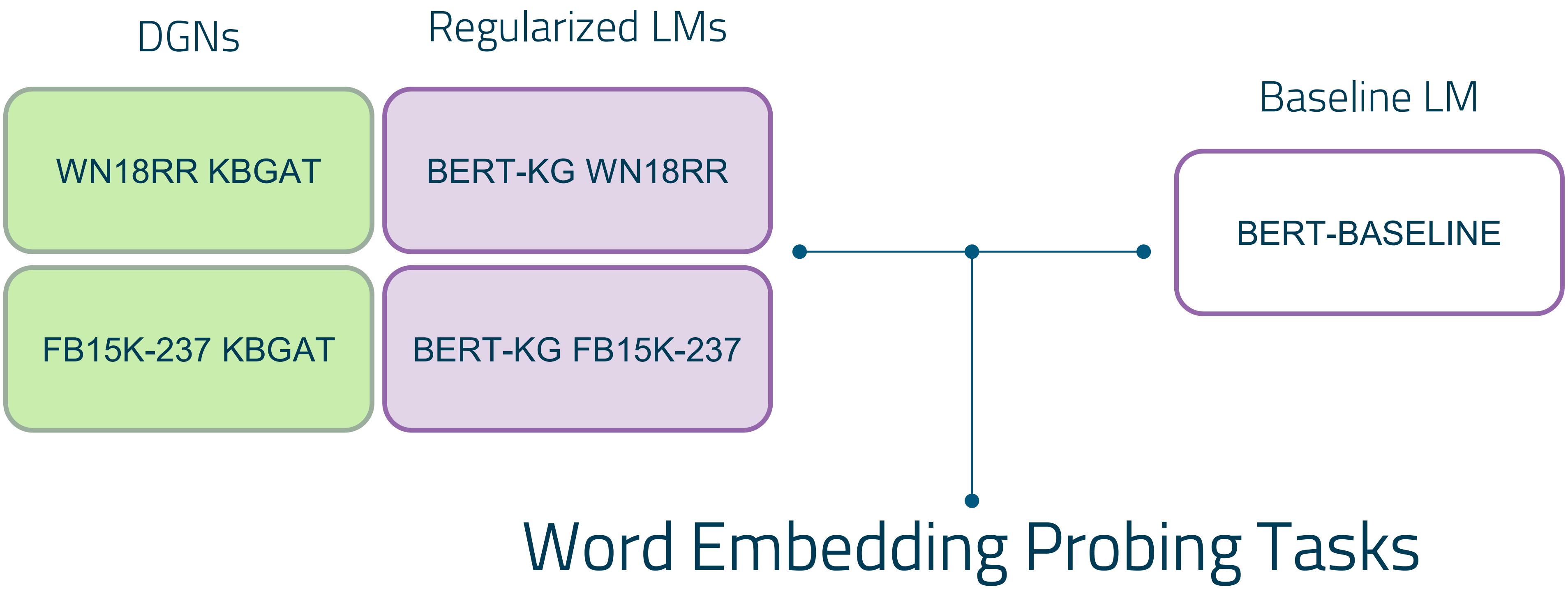
BERT-KG



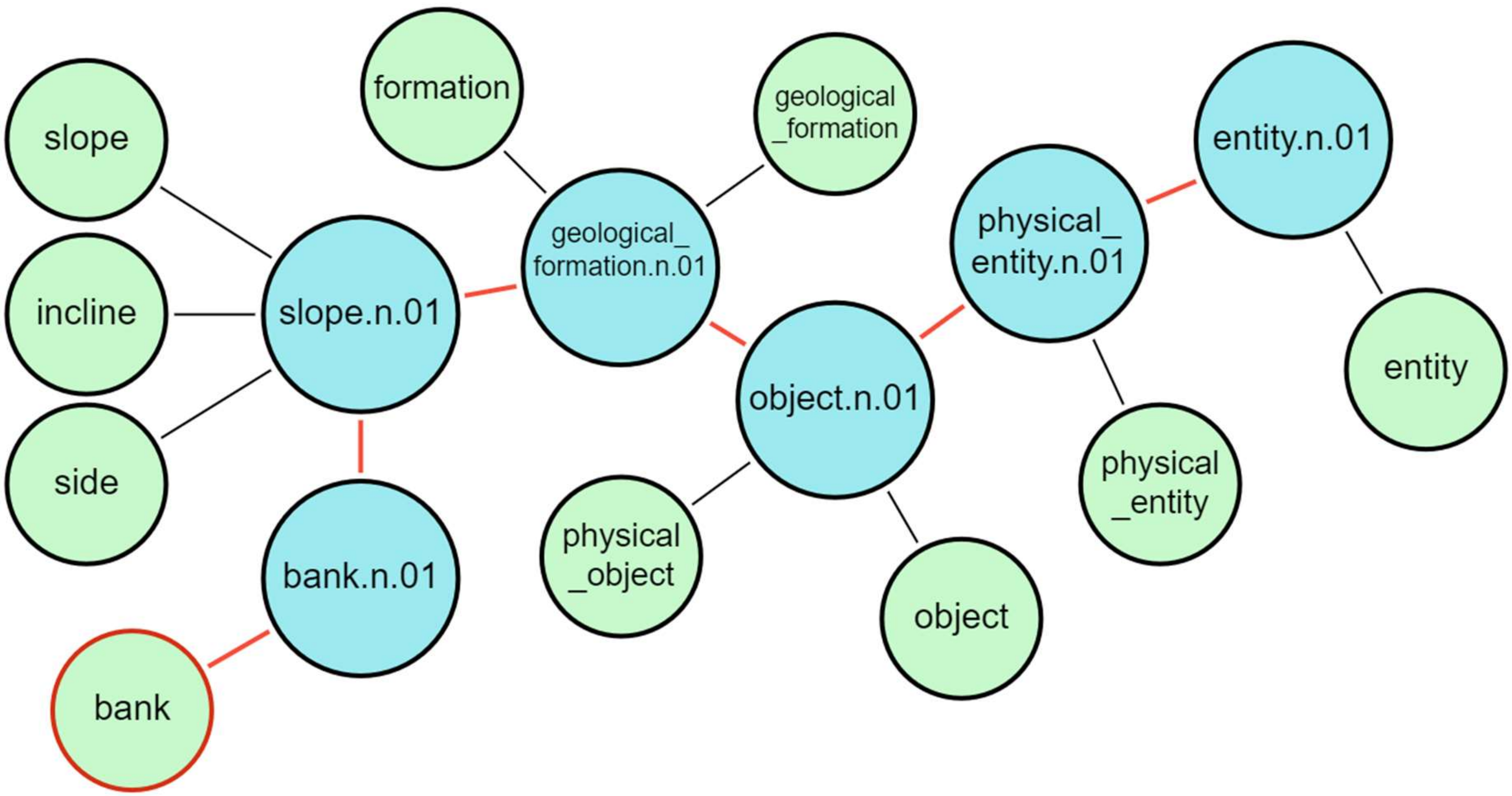
w
768
$\mathbf{x}_i$
512
API
ϕ_{WSD}
WIKIDATA
NLTK
$Encoder_{BERT}$
12 layers
$D_{\mathbf{w}}[k \to v]$
k
$\mathbf{H}_{KBGAT}$
$|\mathcal{V}|$
200
512
$\mathbf{z}_i$
$\mathbf{Z_w}$
768
$\| \|_2^2$
512
$\mathbf{h}_i$
$\mathbf{H_w}$
512
$\mathbf{Z_w A}$
200
200
$\mathcal{L}_R(\mathbf{w})$

EXPERIMENTS

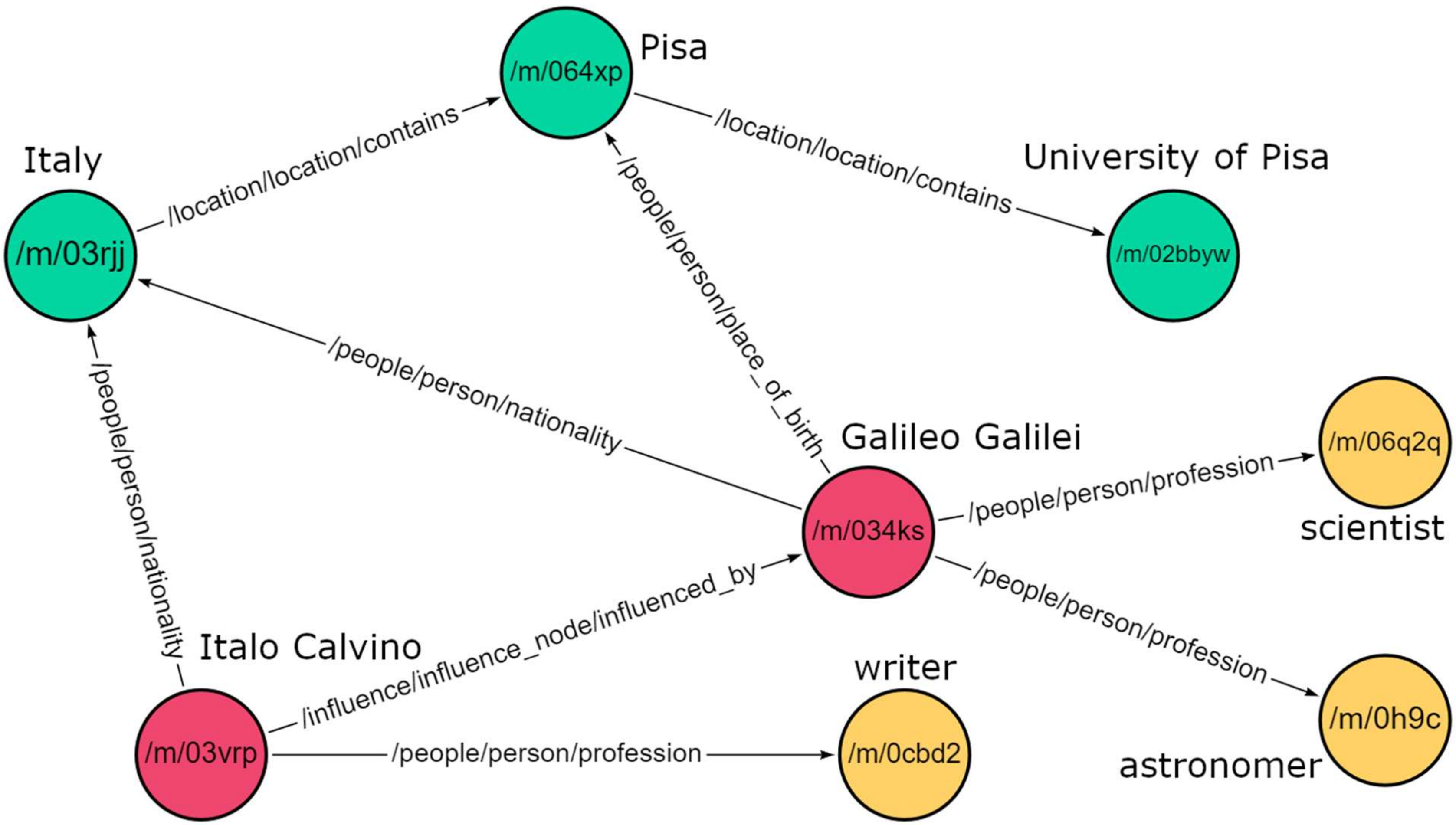
EXPERIMENTAL SETUP



WN18RR DATASET



FB15K-237


Pisa
/m/064xp
Italy
/m/03rjj
/location/location/contains
/location/location/contains
University of Pisa
/m/02bbyw
/people/person/place_of_birth
/people/person/nationality
/people/person/nationality
Galileo Galilei
/m/034ks
/people/person/profession
/m/06q2q
scientist
/people/person/profession
/m/0h9c
astronomer
Italo Calvino
/m/03vrp
/influence/influence_node/influenced_by
/people/person/profession
writer
/m/0cbd2

WORD SIMILARITY PROBING TASK

MEN Dataset (*Bruni et al, 2014*)
3000 rows

Word 1	Word 2	Similarity Score
sun	sunlight	50
automobile	car	50
river	water	49
stair	staircase	49
morning	sunrise	49
feather	truck	1
festival	whisker	1
muscle	tulip	1
bikini	pizza	1
bakery	zebra	0

Regression Task to predict Word Embedding Similarity

- k-Fold Cross Validation k = {10, 20 50, 100}

- Metrics:

$$\mathrm{MSE}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$\mathrm{R2}(\mathbf{y}, \hat{\mathbf{y}}) = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

$$\rho(\mathbf{y}, \hat{\mathbf{y}}) = 1 - \frac{6 \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n(n^2 - 1)}$$

KB COMPLETION PROBING TASK

SQuAD Dataset (*Rajpurkar et al. 2016*)

Q: "Who developed the theory of relativity?"
A: "The theory of relativity was developed by [MASK]"
GT: "Albert Einstein"

Q: "Where was Galileo Galilei born??"
A: "Galileo Galilei was born in [MASK]"
GT: "Pisa"

Querying a language model for factual knowledge as cloze test

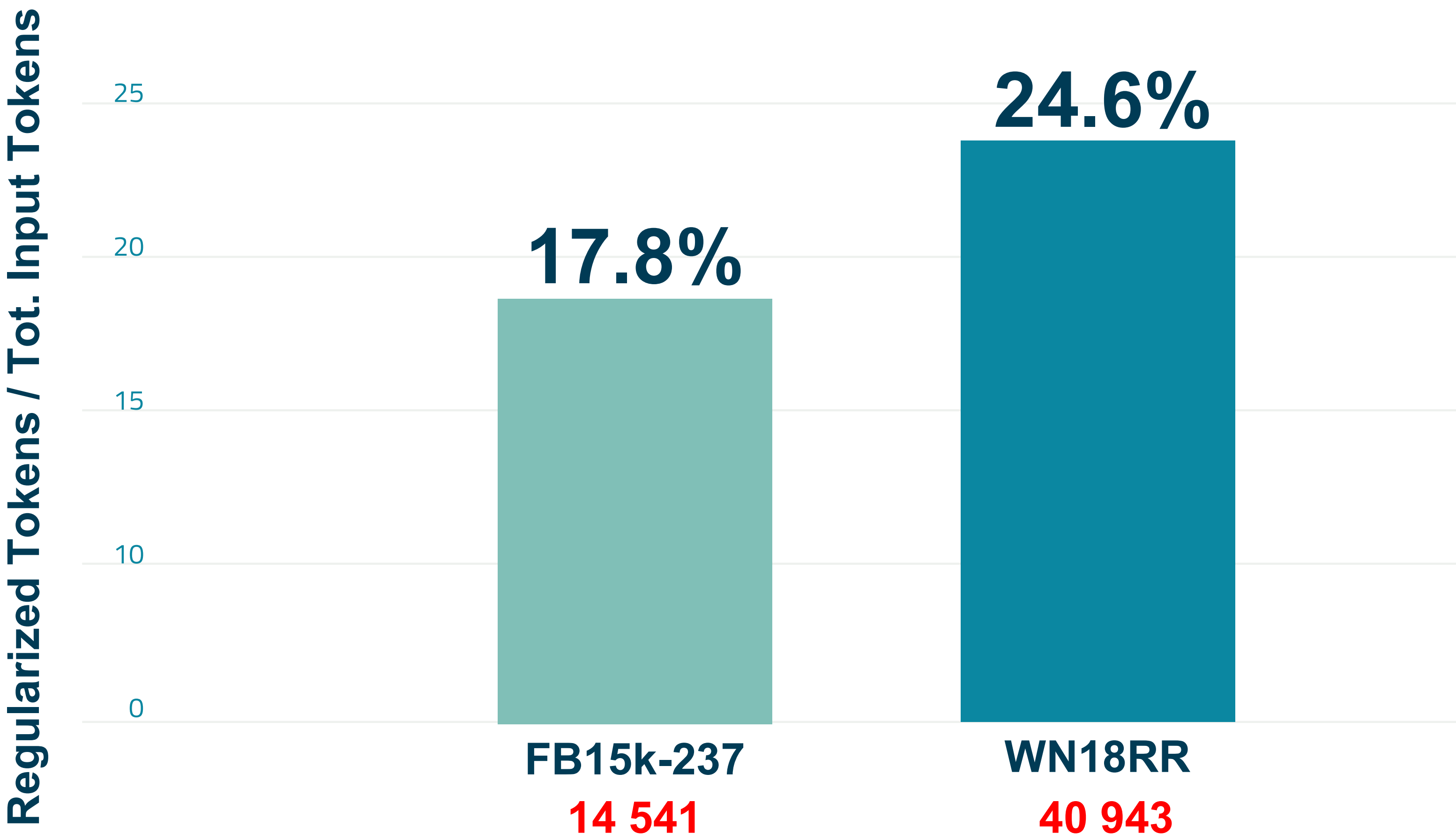
- LAMA Probe (Petroni et al. 2019): 305 question-answer rows from SQuAD
- Metrics:

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\mathrm{rank}(token_*^i)}$$

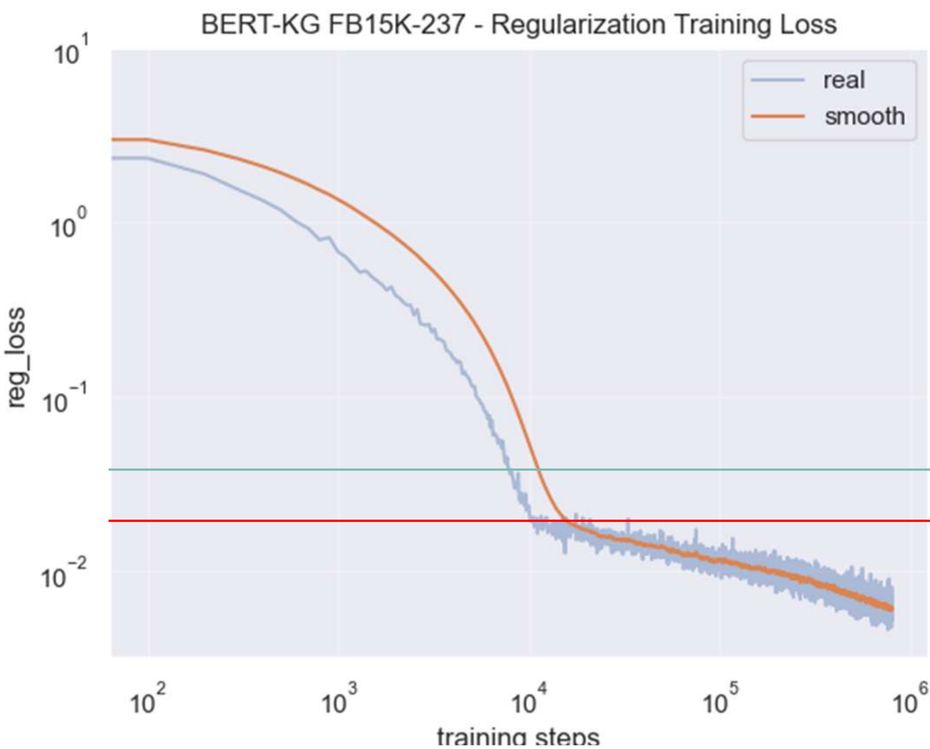
$$\mathrm{P@}K = \frac{1}{|Q|} \sum_{i=1}^{|Q|} H\left(K - \mathrm{rank}\left(token_*^i\right)\right)$$

RESULTS AND ANALYSIS

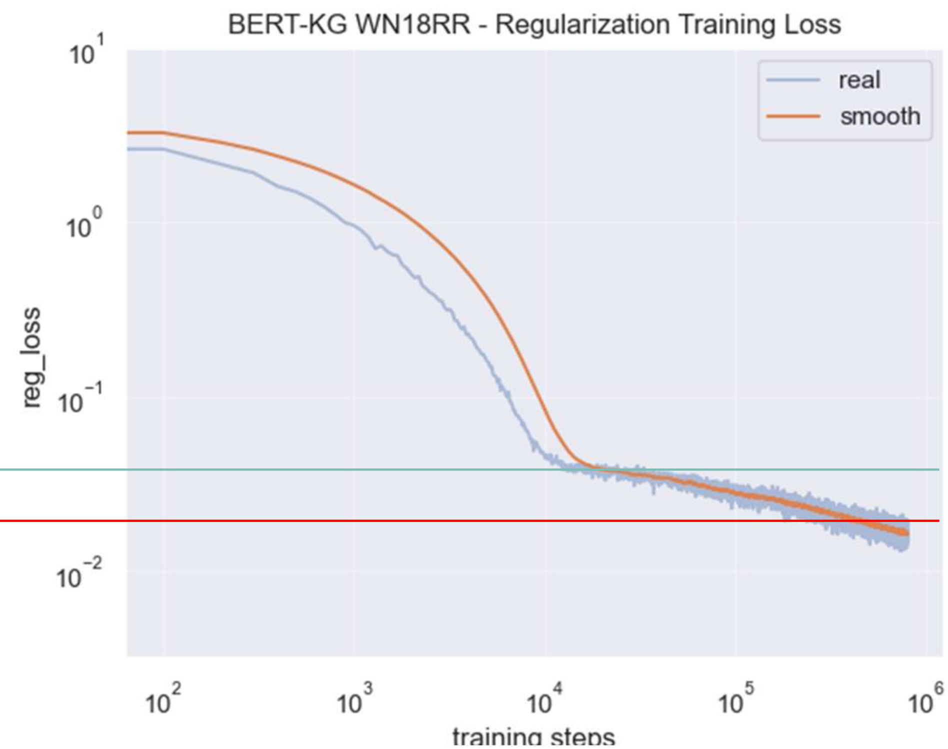
REGULARIZATION RATIO



LMs REGULAIZATION LOSS



FB15k-237

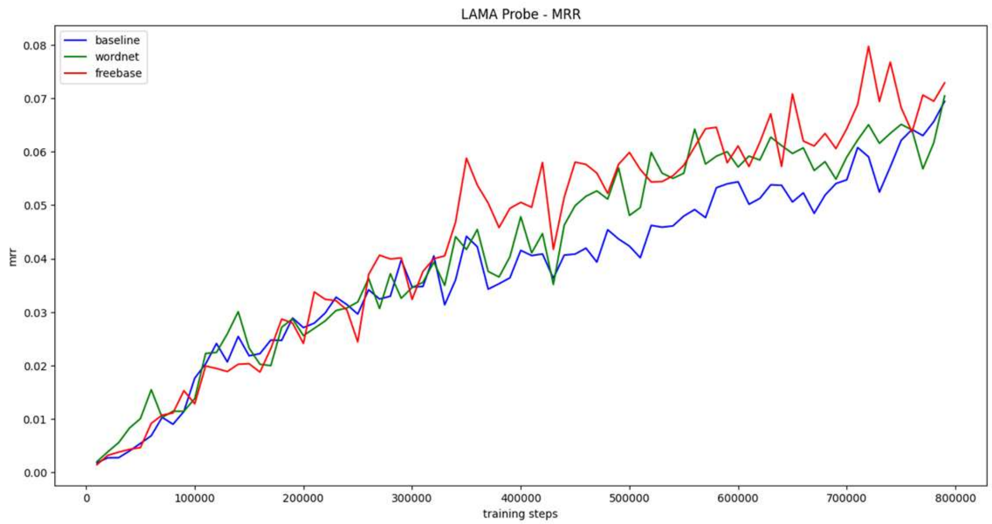


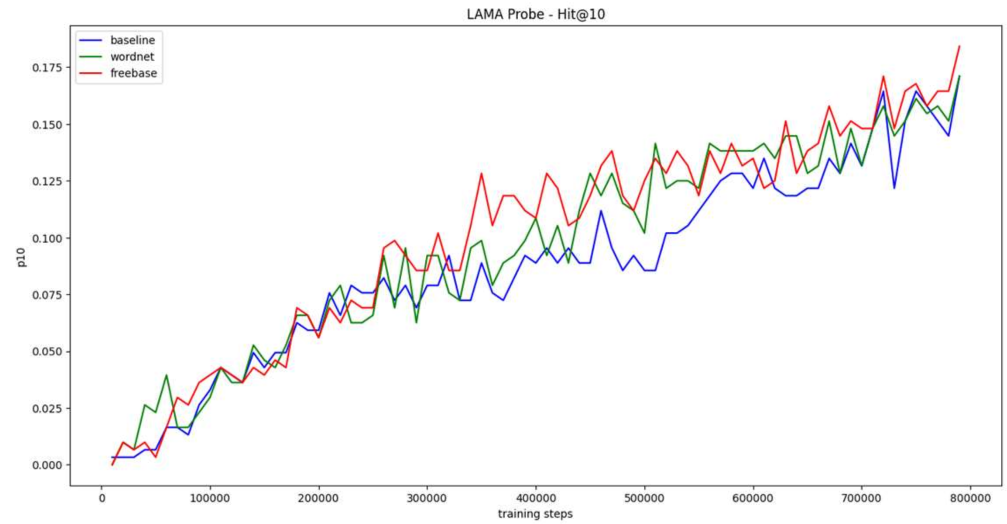
WN18RR

WORD SIMILARITY

Model	MSE		R2		ρ	
20-fold 800k	mean	std	mean	std	mean	std
BERT-BASELINE	0.039	0.004	0.305	0.075	0.601	0.059
BERT-KG WN18RR	**0.036**	0.003	**0.362**	0.076	**0.652**	0.069
BERT-KG FB15K-237	0.039	0.006	0.311	0.112	0.614	0.087
bert-base-uncased	0.017	0.003	0.714	0.079	0.854	0.034

KB COMPLETION





Model	MRR	P@10
BERT-BASELINE	0.070	0.172
BERT-KG WN18RR	0.071	0.174
BERT-KG FB15K-237	**0.074**	**0.184**
bert-base-uncased	0.47	0.25

CONCLUSIONS

1. MLM regularization term was able to transfer part of the symbolic knowledge into the parameters of the BERT language model

2. Graph-driven regularization does not degrade the performance of the language models

3. The regularization approach can be extended to other deep learning models, encoding symbolic knowledge as knowledge graph embeddings and transferring it into multi-domain embeddings

4. The regularization is strongly related to the ontological domain underlying the KG employed in regularization

5. Analyze the influence of KG-regularization on multiple graphs in language modeling

6. Reduce the high number of transformer parameters

THANK YOU FOR YOUR ATTENTION.

Any questions?