Large Language Models for Bulgarian NLP

Nikolay Paev, Kiril Simov, Petya Osenova, Silvia Petrova
IICT, BAS

CLaDA-BG 2024 Conference
26-28 June 2024

Plan of the Talk

- Introduction
- Language Models
- BERT and LlaMa Models
- NLP Tasks for Fine-tuning
- Results
- Conclusions and Future work

BulTreeBank CLaDA BG

Introduction: Aim

- Our main aim is to pretrain several Large Language Models to support different tasks within CLaDA-BG
- We have started with training smaller models – BERT and we have performed several experiments training on different corpora, with different (hyper) parameters and different model size
- Our first application goal is to construct efficient language pipe for Bulgarian
- We have also performed some experiments with respect to generation of pseudo corpora for further training of LLMs

Language Models

- Language modelling:

  everyone in the room was ___

  [listening 20%, talking 15%, …]

- Masked Language modelling:

  they ___ *in silence.*

  [listened 6%, watched 8%, …]

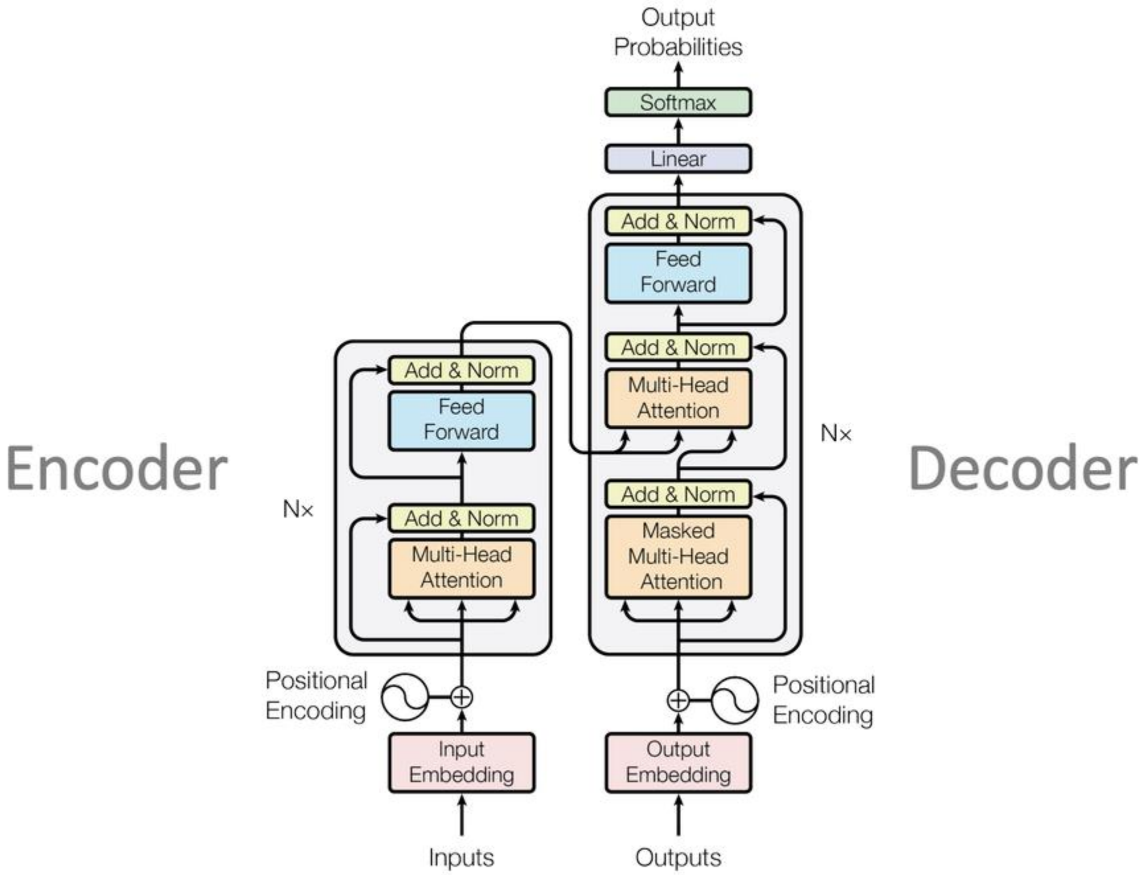
Transformers

“Attention Is All You Need” - (Vaswani et al. 2017)

- encoder
  context => vector representations

- decoder
  context => shifted context

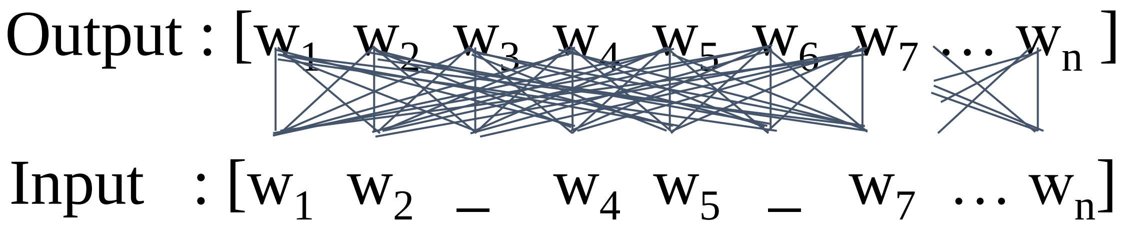
separated as different models



Pre-training tasks

Unsupervised corpora - billions of words

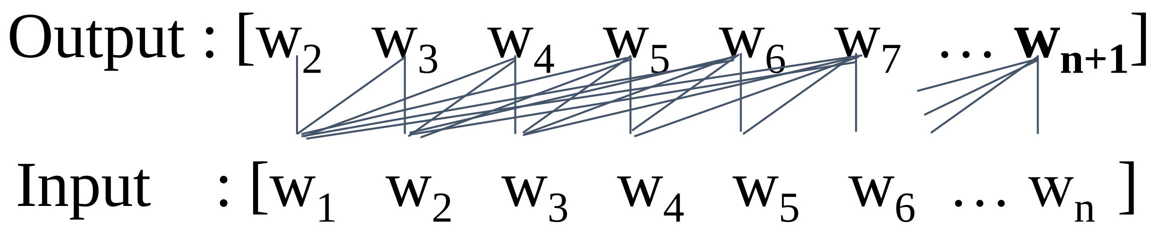
- Masked word prediction
  - encoder models
  - bi-directional attention



Output : [w_1 w_2 w_3 w_4 w_5 w_6 w_7 … w_n]

Input : [w_1 w_2 _ w_4 w_5 _ w_7 … w_n]

- Next word prediction
  - decoder models
  - causal attention



Output : [w_2 w_3 w_4 w_5 w_6 w_7 … $\mathbf{w_{n+1}}$]

Input : [w_1 w_2 w_3 w_4 w_5 w_6 … w_n]

Pre-trained models

Dataset - News articles + Literature

The more diversity in the dataset the better

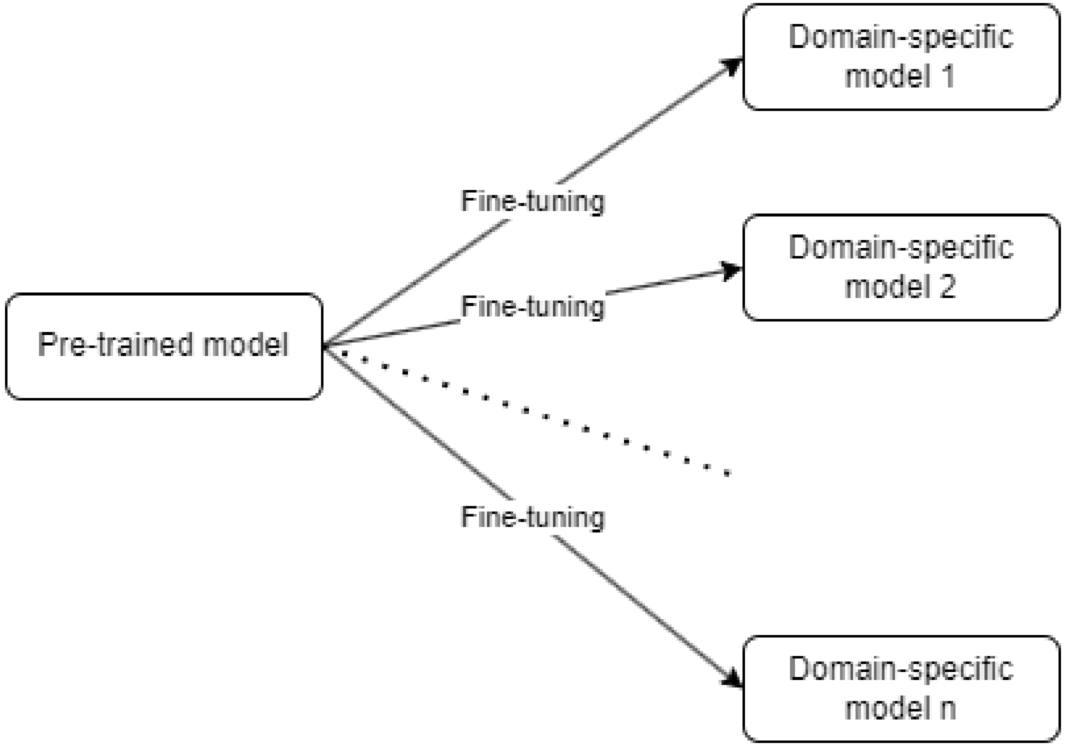
- BERT - base 109M
- BERT - middle 183M
- BERT - large 334M

- LlaMa - small 934M

Fine-tuning

- encoder + classifier layer
  - text classification
  - token classification

- decoder
  - generation in specific domain
    - Question answering
    - Summarization
    - Information extraction



Pre-trained model

Fine-tuning

Fine-tuning

Fine-tuning

Domain-specific model 1

Domain-specific model 2

Domain-specific model n

NLP Tasks for Fine-tuning

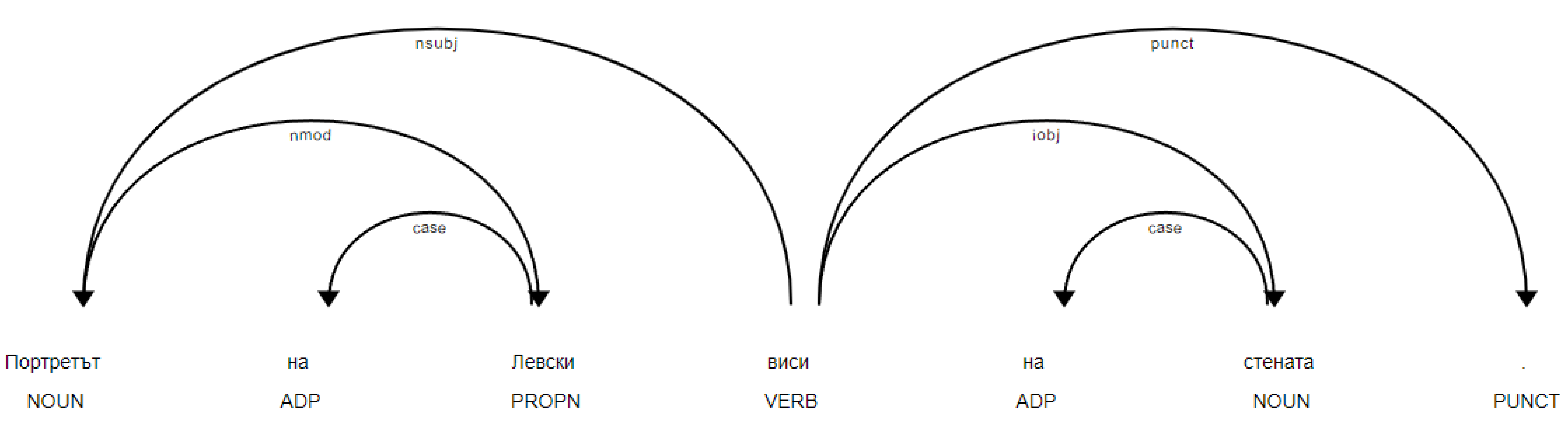
- Part-of-speech Tagging - UPOS, XPOS
- Named Entity Recognition

word	**Портретът**	**на**	**Левски**	**виси**	**на**	**стената**	**.**
UPOS	NOUN	ADP	PROPN	VERB	ADP	NOUN	PUNCT
NER	O	O	B-PER	O	O	O	O

word	**Левски**	**отново**	**се**	**класира**	**за**	**Лига**	**Европа**	**.**
UPOS	PROPN	ADV	PRON	VERB	ADP	NOUN	PROPN	PUNCT
NER	B-ORG	O	O	O	O	B-OTH	I-OTH	O

NLP tasks for Fine-tuning

- UD parsing - linking, classification



Results

Bigger models and bigger pre-training corpora* - better fine-tuning results

Pre-trained model	UPOS accuracy	XPOS accuracy	NER bs micro F1	NER np micro F1	UD combined accuracy
BERT-base Lit + Articles 2020-2021	99.0%	97.7%	98.5%	83.2%	90.0%
BERT-middle Lit + Articles 2020-2021	99.2%	98.0%	99.5%	85.6%	91.0%
BERT-base Lit + All Articles	99.1%	97.9%	99.9%	85.7%	91.1%
BERT-large Lit + All Articles	99.4%	98.2%	99.9%	85.7%	92.1%

Decoder Fine-tuning

- Text denoising - creating pseudo-corpora

In:	*велик цел заслужавам всякакъв жертва .*
Out:	*великите цели заслужават всяка жертва .*

- Question Answering

In:	*Кога е роден Христо Ботев ?*
Out:	*Христо Ботев е роден на 6 януари 1848 г . в Калофер .*

- Generating definitions

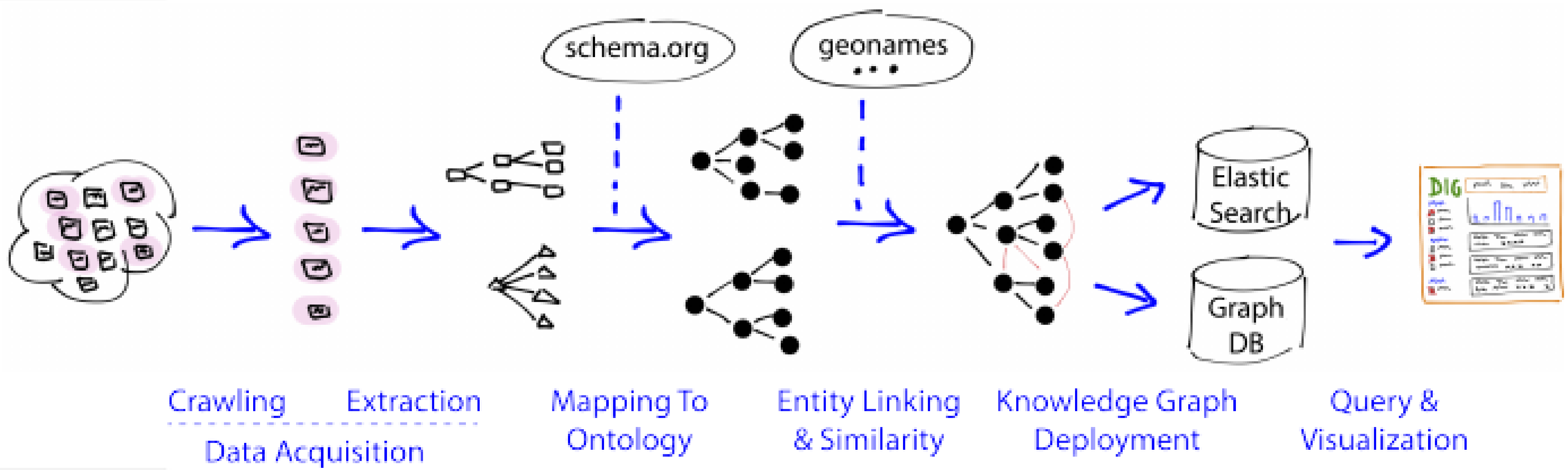
In:	*телескоп*
Out:	*оптичен уред за наблюдаване на небесните тела .*

Application in Humanities & Social Sciences

- Indexing of documents with Named Entities

- Extracting knowledge from text

LLM in knowledge extraction



Conclusion

- Pre-trained language models for Bulgarian
- Achieved best results on different NLP tasks for Bulgarian
- Created a language annotation pipe and API

Future Plans

- Opening the API to the public
- Uploading the weights of the best models to HuggingFace
- Gathering and cleaning more data and pre-training larger models
- Experiments with representation of language resources as text for pre-training
- Other tasks for the domain of humanities and social sciences

References

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is All You Need." In Advances in Neural Information Processing Systems, 5998-6008, 2017.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, Alexander M. Rush. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. https://arxiv.org/abs/1910.03771