DOMAIN GENERALIZATION IN VIETNAMESE DEPENDENCY PARSING

A NOVEL BENCHMARK AND DOMAIN GAP ANALYSIS

Vinh-Hien D. Huynh, Chau-Anh Le,

Chau M. Truong, Y Thien Huynh and Quy T. Nguyen

University of Information Technology, Ho Chi Minh City, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam

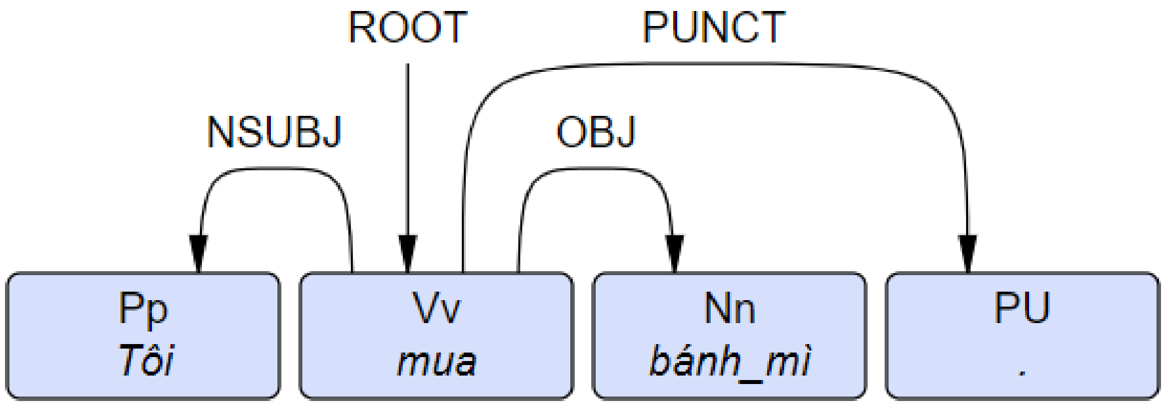
ĐẠI HỌC QUỐC GIA
THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN

Introduction: Dependency Parsing

Dependency Parsing: build a directed tree by Grammatical relations

Tôi mua bánh mì.
(I buy some bread.)



Introduction: Domain Gap

Domain Gap: The difference between training and testing domain(s)

Training:

- *I love you so much.*
- *You are the light of my life.*
- *I'm crazy about you.*

Testing:

- *Em an com chua.*

Problem

- There is a limited number of research inside domain generalization on Vietnamese dependency parsing.
- One of the keys reason is the lack of a compatible treebank.

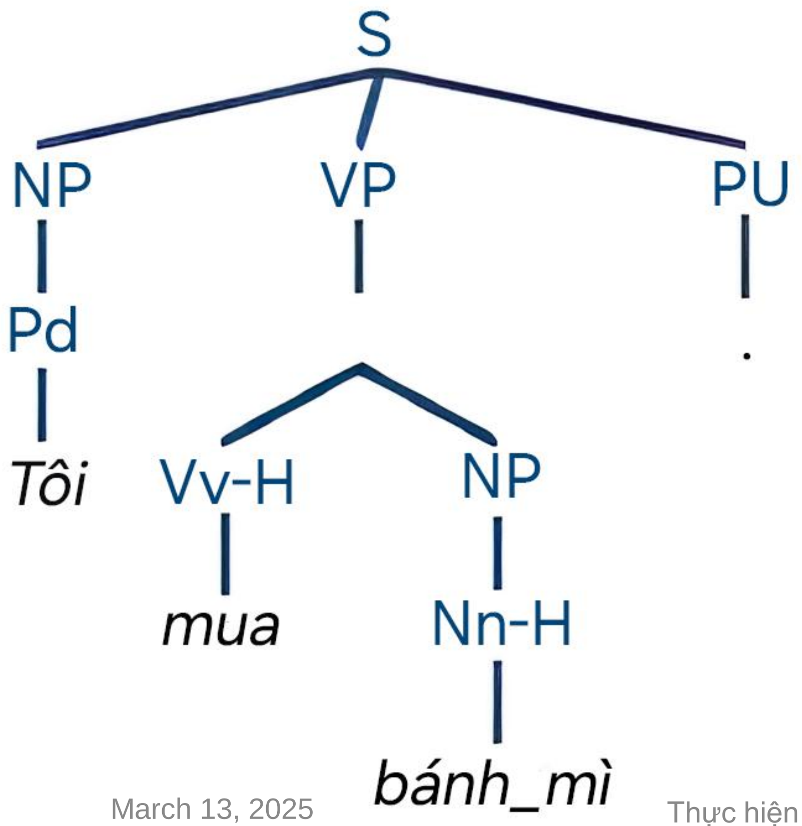
Our contribution

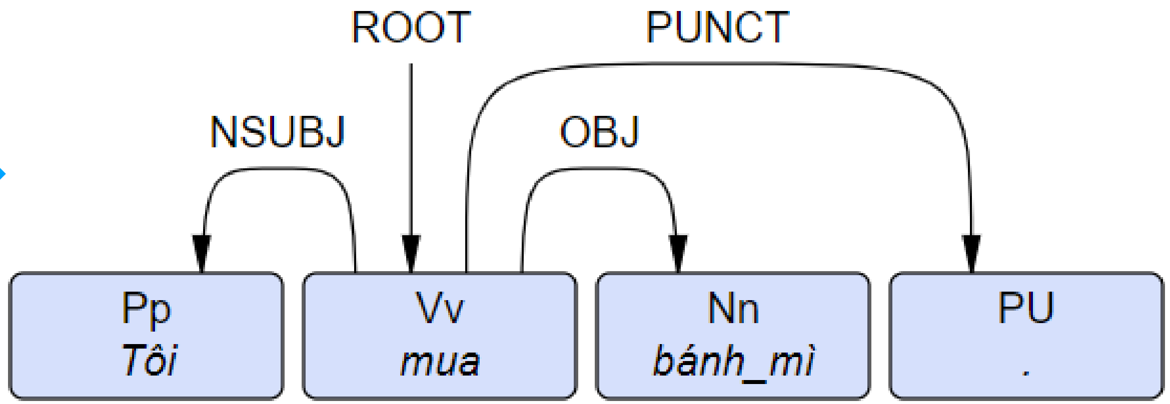
In this research, we release:

- DGDT, a Vietnamese dependency treebank that available in multiple domains, which can accommodate cross-domain setups.
- DGDTMark, a benchmark suite available in different cross-domain scenarios.

Dataset construction

- To keep balance between cost and data quality, we adopt a converter to transform an existing constituency treebank.





Dataset construction

- We follow the converter released by Truong et al. (1) and the constituency treebank of Nguyen et al. (2) as they fit our requirements.

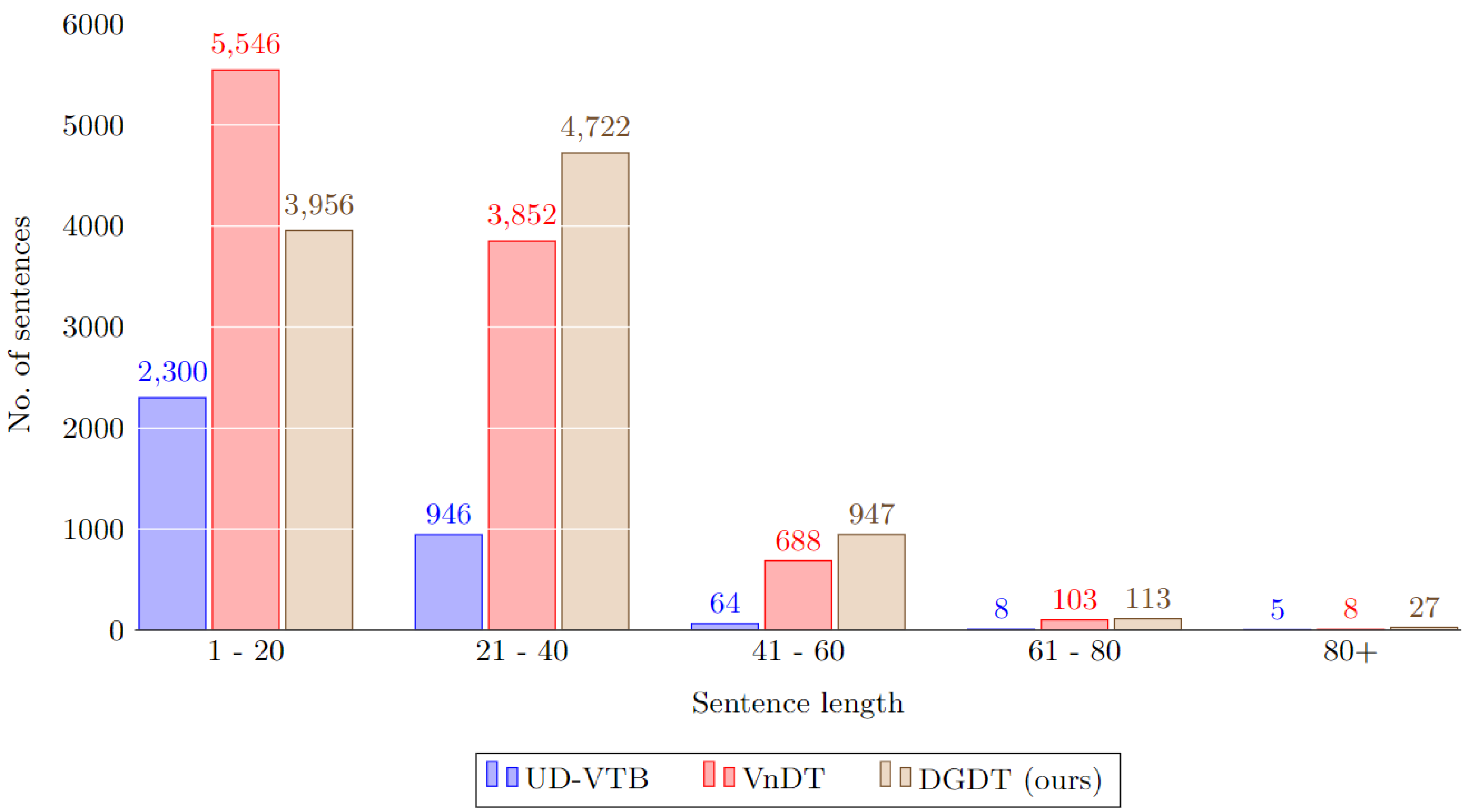
- Our dataset contains 9,765 sentences in 14 newspaper topics, crawled from the Thanh Nien Newspaper.
- We treat each topic as a single domain.

(1) C. M. Truong, T. V. Pham, M. N. Phan, N. D. T. Le, T. V. Nguyen and Q. T. Nguyen, "Converting a constituency treebank to dependency treebank for Vietnamese," 2022 RIVF International Conference on Computing and Communication Technologies (RIVF), Ho Chi Minh City, Vietnam, 2022, pp. 256-261
(2) Nguyen, Q.T., Miyao, Y., Le, H.T.T. et al. Ensuring annotation consistency and accuracy for Vietnamese treebank. Lang Resources & Evaluation 52, 269–315 (2018).

Dataset Statistics

Set	Domain	Number of Sentences
Train	Education	844
	Health	725
	Law	610
	Life_of_youth	635
	Military	690
	Politics_Society	712
	Science	692
	Sports	697
	Travel	540
	World	645
Dev	Entertainment	708
	Information_Technology	714
Test	Economic	725
	Life	828
	Total	9765

Dataset Statistics



The DGDTMark

To evaluate the effects of domain gap into the task, we released this benchmark suite, includes 4 scenarios:

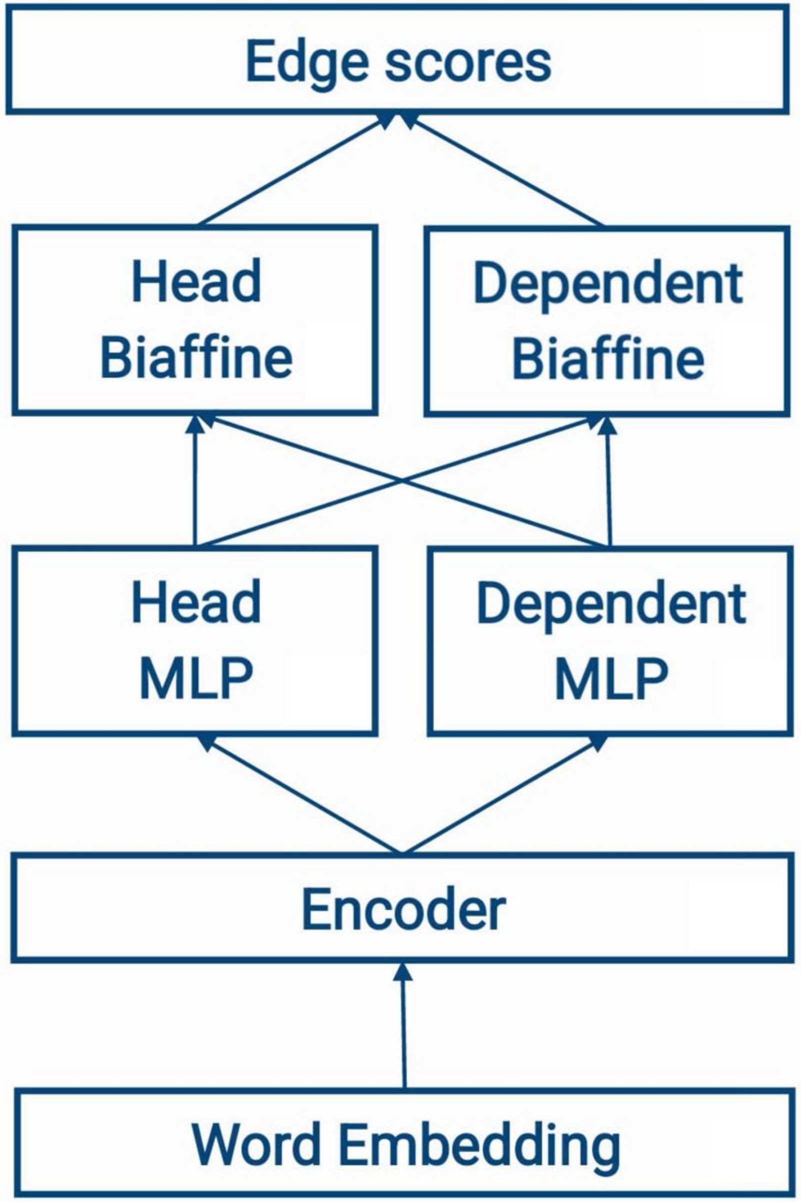
1. *in-domain:* split each domain with ratio 8:1:1 and merge the respective parts from all domains to build train/dev/test sets.
2. *domain-k-fold:* let each domain be the dev and test set, while the rest of treebank plays the role of train set.
3. *domain-generalization:* arrange domains exclusively to one of three train/dev/test sets.
4. *dataset-generalization: replace the test set of DGDT in the third scenario with test set of the NIIVTB_DT-1 (1) treebank.*

(1): first reference on the 7th slide

Models

- We use the Deep Biaffine Attention Model as our baseline for DGDTMark.
- We run our experiments on the implementation of Zhang et al. (3) with default hyperparameters.
- The original LSTM is replaced by transformer-based encoder including PhoBERT and XLM-RoBERTa (XLM-R).

(3) https://github.com/yzhangcs/parser

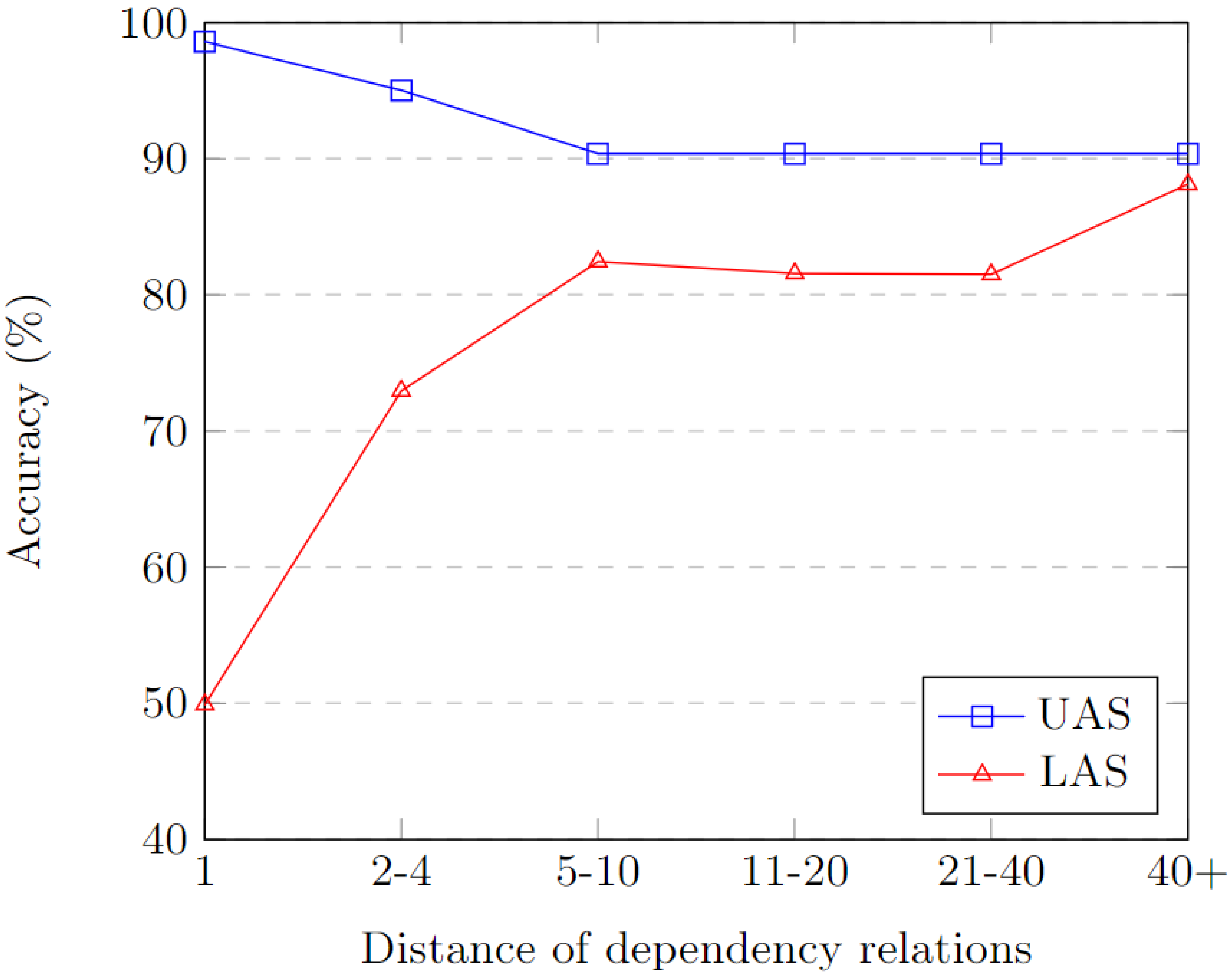


Experiment Results

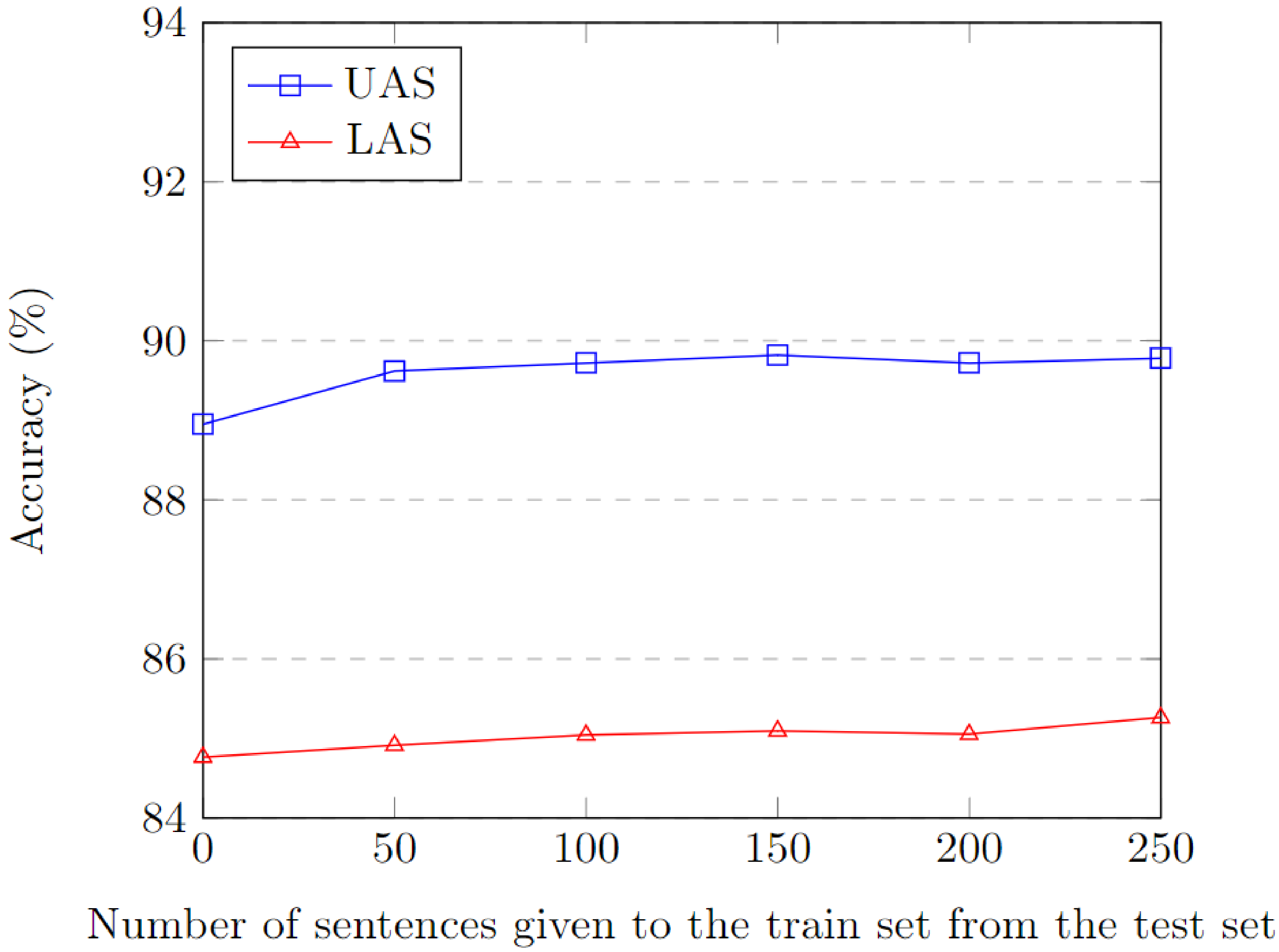
Encoder	in-domain	domain-k-fold	domain-generalization	dataset-generalization
PhoBERT	87.83	87.01	84.76	82.74
XLM-R	85.89	84.99	82.88	80.97

- Both encoders are evaluated with their *base* version.
- The evaluation metric is LAS.

How do the models capture long dependencies?



Can data from testing domains help the models?



Conclusion

- We introduce DGDT treebank and DGDTMark benchmark suite.
- Current models have performed well in this parsing task (87.83% LAS), but have still faced difficulties when handling the domain gap (>5% decrease of LAS).

THANK YOU

Q&A