Probing LLMs for Multilingual Discourse Generalization Through a Unified Label Set

Florian Eichin*, Yang Janet Liu*, Barbara Plank, and Michael A. Hedderich
MaiNLP, Center for Information and Language Processing, LMU Munich, Germany
Munich Center for Machine Learning (MCML)

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

MaiNLP CIS

MCML Munich Center for Machine Learning

✤ Introduction & Motivation

Many approaches to CL / NLP focus on sentence-by-sentence analyses, but there are many research questions which cannot be answered without considering **sentences in the context of a larger discourse**.

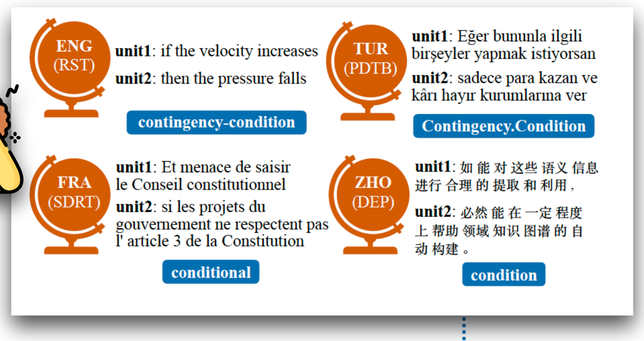
A growing body of research has explored the extent to which PLMs and LLMs encode linguistic representations and exhibit generalizable abstraction but is **limited to morphology, syntax, and semantics**.

RQ1: Do LLMs encode discourse information?
RQ2: Do LLM discourse representations generalize across languages?

- ✤ model size ***alone*** does not lead to discourse probing success; instead, **multilingual training**, **dataset composition**, and **language-specific factors** play significant roles.
- ✤ discourse representations are best aligned across languages in the ***intermediate*** layers, with later layers refining these representations for specific relation types.

want more details? code and paper:

Problem: **No consensus** of **one single set of relations** due to different perspectives on discourse relations

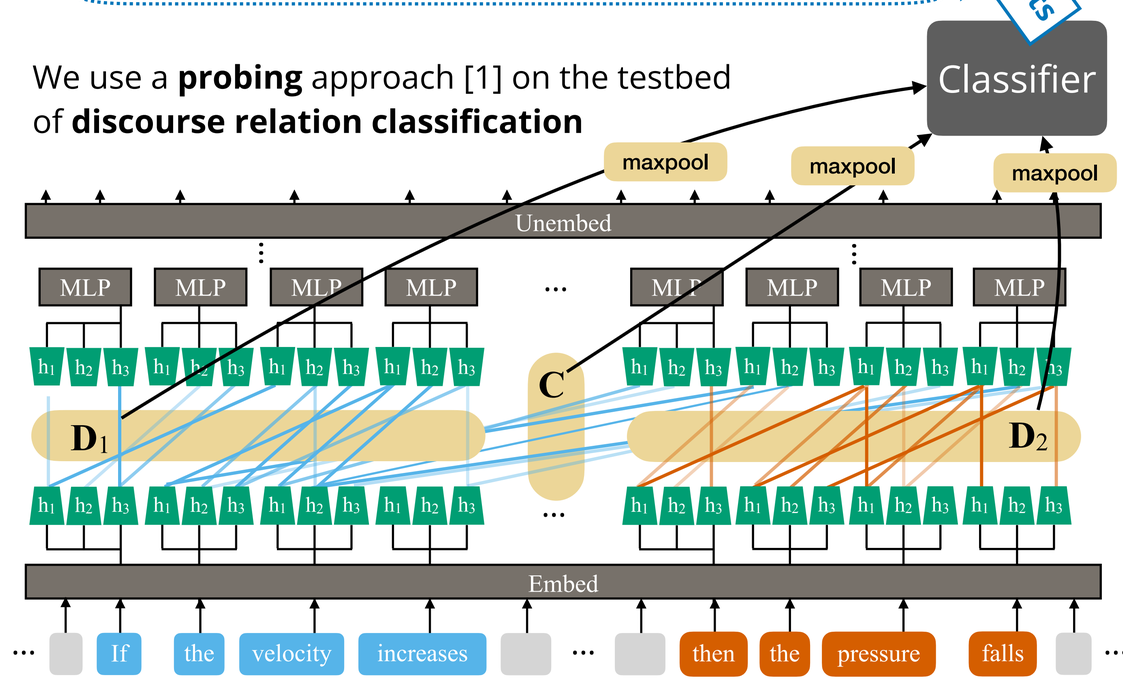


Solution: A unified label set!

structuring	
temporal	
thematic	attribution, comparison, elaboration, framing, mode, reformulation, adversative
causal-argumentative	causal, contingency, enablement, explanation, evaluation
topic-management	topic-change, topic-comment, topic-adjustment

predicts

We use a **probing** approach [1] on the testbed of **discourse relation classification**



✤ Experimental Setup

data: DISRPT 2023 [2]

- ✤ 13 languages from five language families
- ✤ 224,281 discourse relations from 23 corpora across four frameworks

23 decoder-only LLMs

Aya, BLOOM, Phi-4

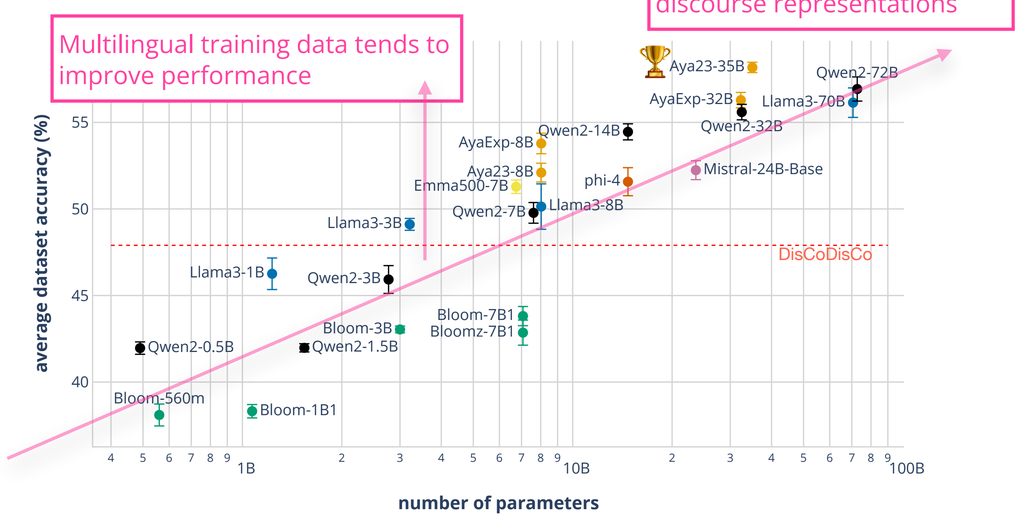
DisCoDisCo (ref sys, [3]): trained in the same, generalized setting across all languages

✤ Results & Findings

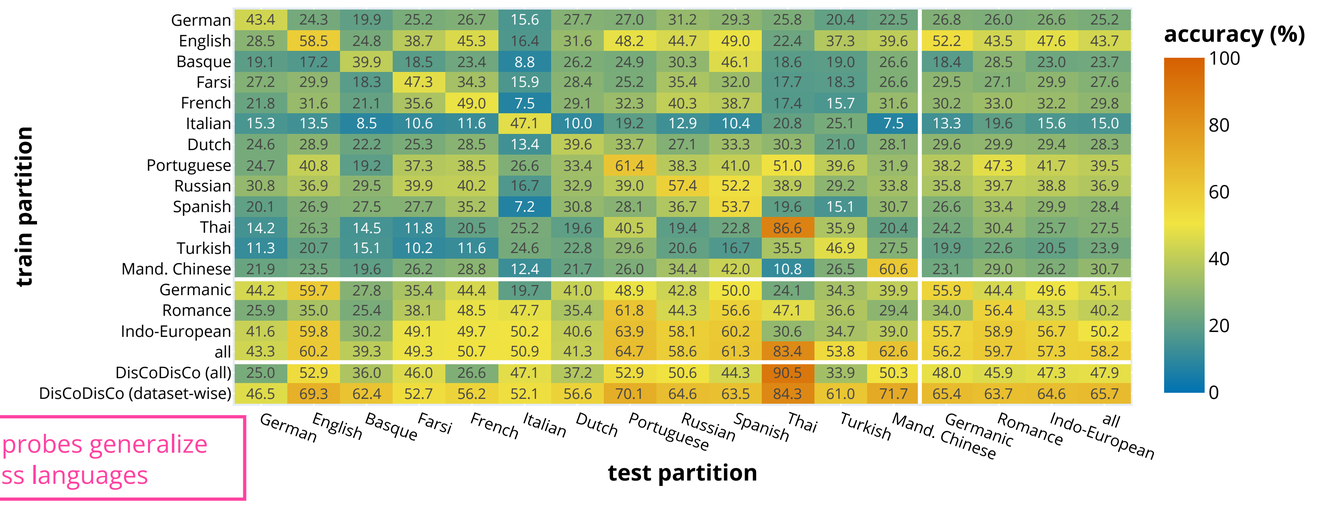
1) overall performance

Larger models have better discourse representations

Multilingual training data tends to improve performance

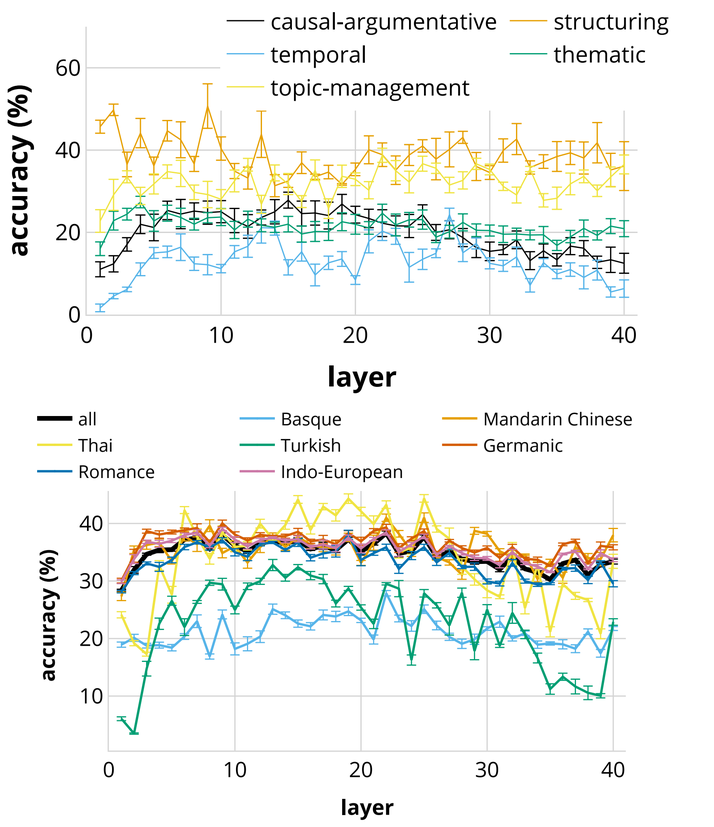


2) by language/language-group performance of Aya-23-35B-probe



Our probes generalize across languages

3) layer-wise performance (Aya-23-35B)



unit1: Respondents were allowed to choose one response from 11 categories.

unit2: For the present analysis , these responses were recoded into nine mutually exclusive categories → elaboration

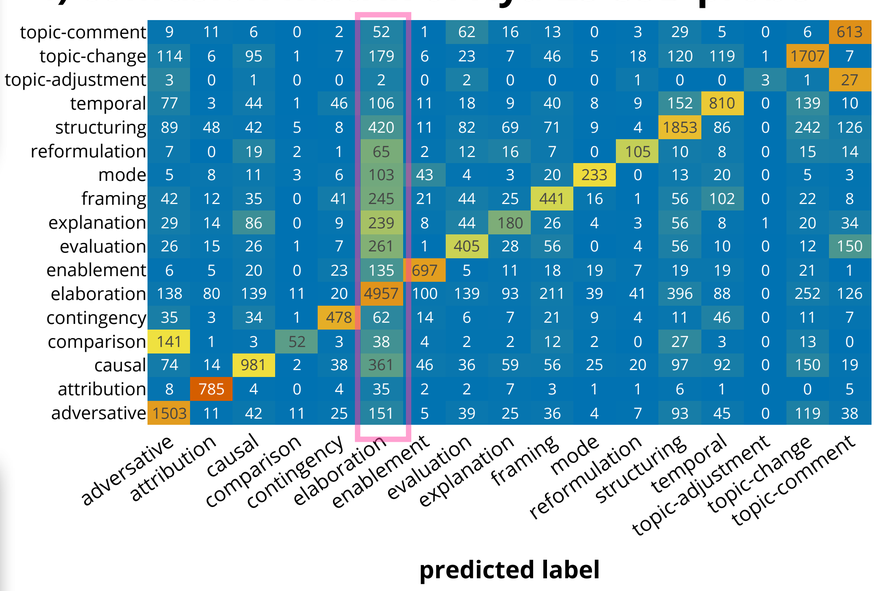
unit1: I was chosen to attend a unique Pilot Training Program,

unit2: The Air Force only selected 180 Officers a year from the Academy , Squadron Officer School and ROTC to attend → framing

unit1: Having mice in your home is both a nuisance and a health hazard. → explanation

unit2: They can cause property damage

4) confusion matrix of Aya-23-35B-probe



[1] Guillaume Alain and Yoshua Bengio. 2018. Understanding intermediate layers using linear classifier probes. *Preprint*, arXiv:1610.01644.
[2] Chloé Braud, Amir Zeldes, Laura Rivière, Yang Janet Liu, Philippe Muller, Damien Sileo, and Tatsuya Aoyama. 2024. DISRPT: A Multilingual, Multi-domain, Cross-framework Benchmark for Discourse Processing. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4990–5005, Torino, Italia. ELRA and ICCL.
[3] Luke Gessler, Shabnam Behzad, Yang Janet Liu, Siyao Peng, Yilun Zhu, and Amir Zeldes. 2021. DisCoDisCo at the DISRPT2021 Shared Task: A System for Discourse Segmentation, Classification, and Connective Detection. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 51–62, Punta Cana, Dominican Republic. Association for Computational Linguistics.