End-to-end Ontology Learning with Large Language Models

Andy Lo, Albert Q. Jiang, Wenda Li, Mateja Jamnik
cyal4@cam.ac.uk, qj213@cam.ac.uk, wenda.li@ed.ac.uk, mj201@cam.ac.uk

Paper
Code

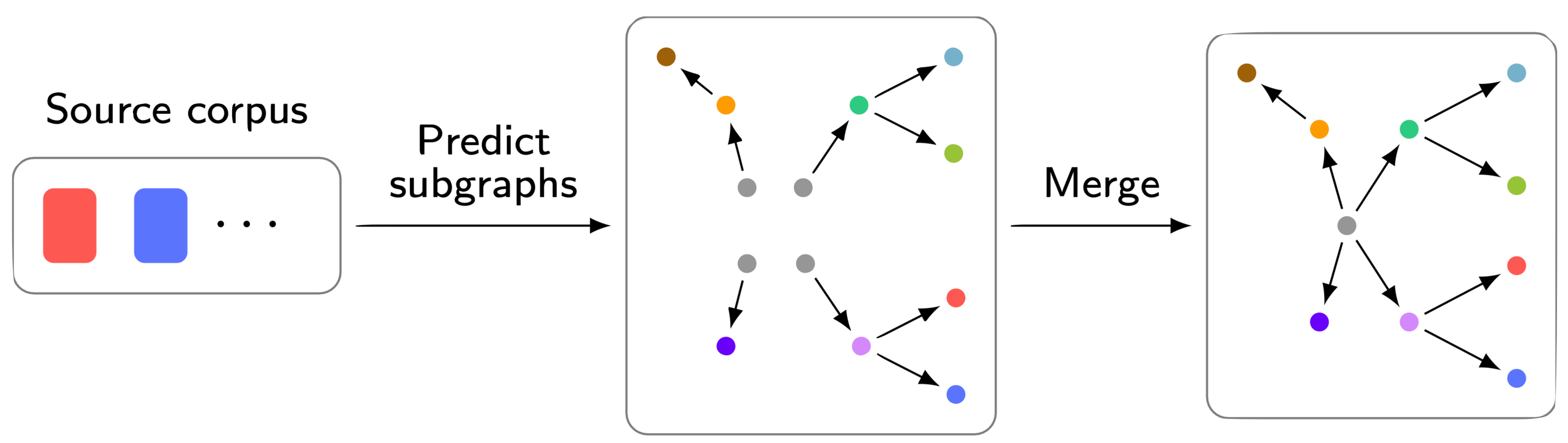
UNIVERSITY OF CAMBRIDGE
THE UNIVERSITY of EDINBURGH

Open problem in OL: Restricted methods and evaluation

Definition:
- An **ontology** is a formal and structural way of representing domain-specific concepts and their relations. Its basic form is a **graph**.

Current method & problems:
- Subtasks composition: First predict **concepts (nodes)**, then **relations (links)**.
- Lacks context between subtasks → **limited performance**.
- Link prediction is O(n²) → **not scalable**.
- Evaluation is done per subtask → **output is not tested directly**.

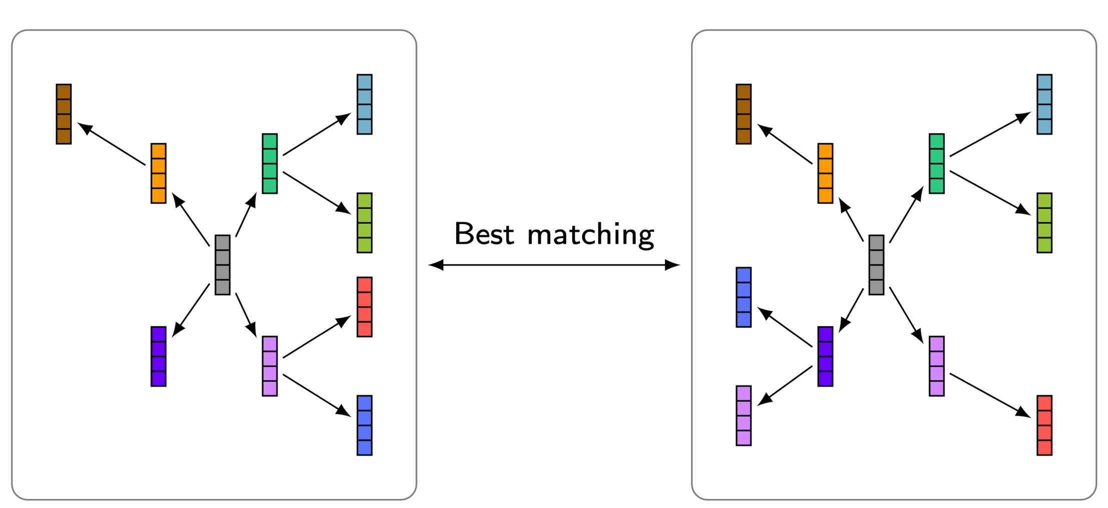
Our solution:
- **Predict subgraphs** and merge by heuristics.
- Measure graph **similarity to ground truth** for evaluation.



New evaluation strategy and metrics

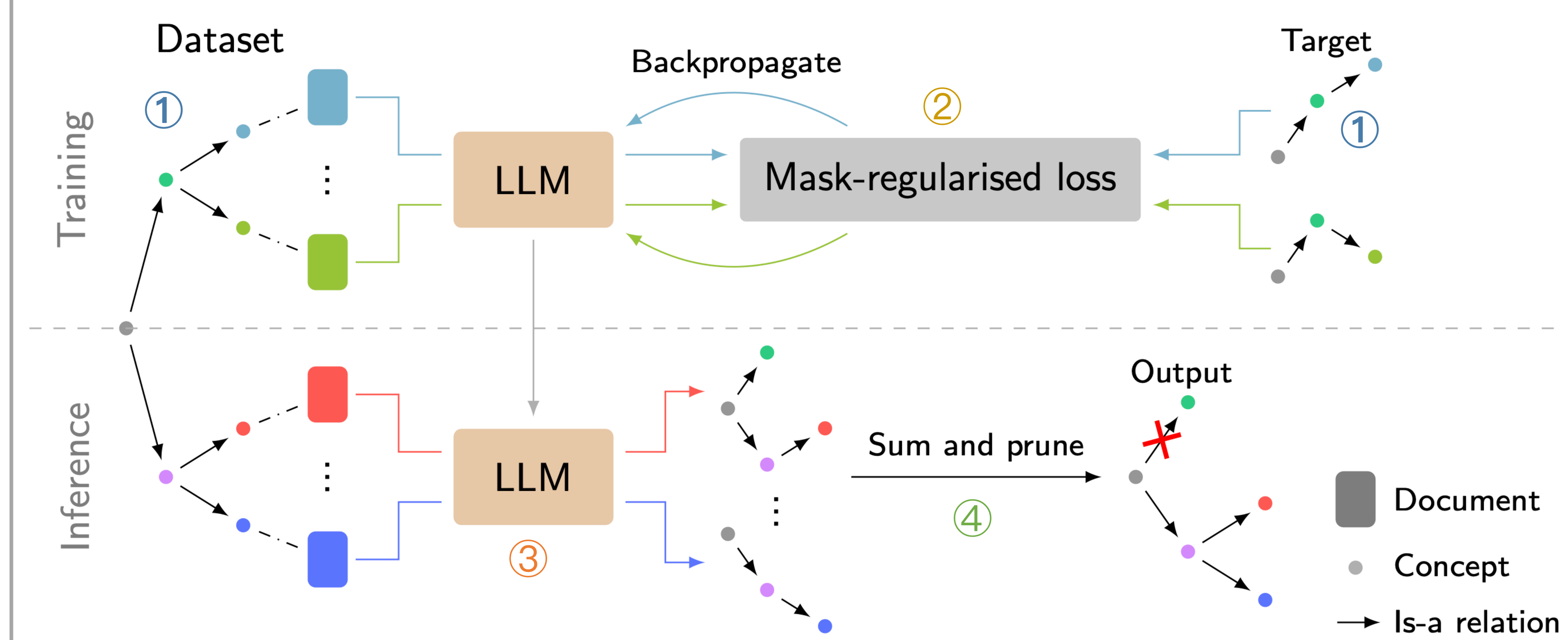
We test the output quality by **golden-standard evaluation**. This requires robust measures of **heterogeneous graph similarity**, but existing methods rely on literal text comparison which is **sensitive to spelling**, **capitalisation**, **choice of word**, etc.

Core principle:
- **Strong emphasis on semantics** over syntax by using pretrained text embeddings.
- Suitable for comparing **arbitrary labelled graphs**.

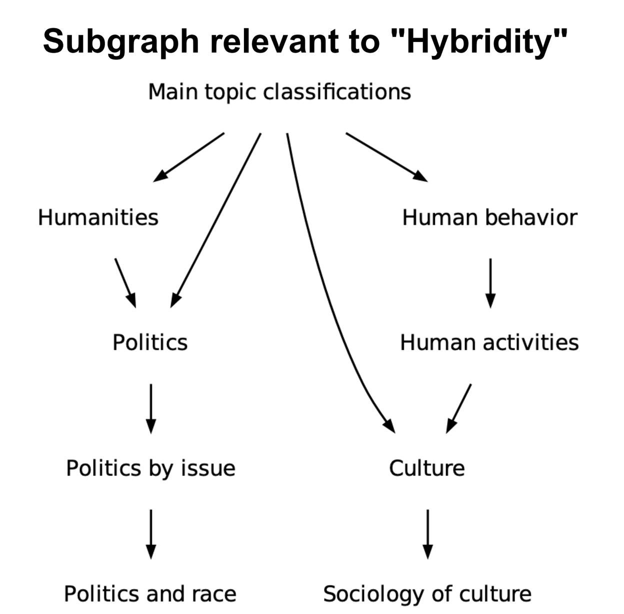


Our new metrics:
- **Fuzzy F1**: Fuzzy edge equality testing for intersection between the graphs.
- **Continuous F1**: Best-scoring edge matching.
- **Graph F1**: Best-scoring node matching.
- **Motif Distance**: Purely structural metric by comparing *network motif* counts.

OLLM: When OL meets LLMs



① Data preparation:
- Split the source ontology into train, validation and test ontologies.
- Given a document and its directly related concepts (leaves), find all short **paths from the root**. Collect all nodes and edges in such paths to form the **target subgraph**.
- Linearise into **text as a collection of paths**.



② Finetuning with custom regulariser:
- By nature of tree-like structure, high-level concepts appear orders of magnitude more often than leaf concepts. **Sampling is highly unbalanced**.
- Naive finetuning results in **overfitting high-level concepts** while **underfitting low-level concepts**.
- To ensure balanced learning, we **mask the loss per concept** with probability proportional to its occurrence frequency.

Only train on highlighted words

<s>[INST] Title: Hybridity
Hybridity, in its most basic sense ... [/INST]
Main topic classifications -> Human behavior -> Human activities -> Culture -> Sociology of culture
Main topic classifications -> Humanities -> Politics -> Politics by issue -> Politics and race
Main topic classifications -> Politics -> Politics by issue -> Politics and race
Main topic classifications -> Culture -> Sociology of culture</s>

③ Model inference:
- Each document-subgraph prediction is independent, allowing us to easily **parallelise inference**.

④ Merging subgraphs:
- Prune edges by simple rules and thresholding. We tune the threshold on the validation set.
- Unlinked nodes are discarded.

Result: Better models, more robust metrics

- **OLLM outperforms all methods** except on Literal F1 and Motif Distance.
- Memorising the training set already gives very strong performance on Literal F1 and Motif Distance → They are strongly **biased towards overfitting!**

Dataset	Method	Literal F1 ↑	Fuzzy F1 ↑	Cont. F1 ↑	Graph F1 ↑	Motif Dist. ↓
Wikipedia	Memorisation	**0.134**	0.837	0.314	0.419	0.063
	Hearst	0.003	0.538	0.350	0.544	0.163
	Rebel	0.004	0.624	0.356	0.072	0.132
	Zero-shot	0.007	0.871	0.455	0.639	0.341
	One-shot	0.031	0.888	0.477	0.610	0.314
	Three-shot	0.031	0.880	0.475	0.622	0.354
	Finetune	0.124	0.884	0.470	0.588	**0.050**
	OLLM	0.093	**0.915**	**0.500**	**0.644**	0.080
arXiv	Memorisation	0.000	0.207	0.257	0.525	**0.037**
	Hearst	0.000	0.000	0.151	0.553	0.098
	Rebel	0.000	0.060	0.281	0.546	0.088
	Zero-shot	0.025	0.450	0.237	0.414	0.145
	One-shot	**0.072**	0.460	0.290	0.433	0.293
	Three-shot	0.051	0.405	0.212	0.385	0.124
	Finetune (transfer)	0.000	0.440	0.225	0.441	0.148
	OLLM (transfer)	0.040	**0.570**	**0.357**	**0.633**	0.097

Table 1: Evaluation metrics of OLLM and baselines on Wikipedia and arXiv. OLLM performs particularly well in modelling semantics, and remains competitive syntactically and structurally.

Example: Generation and evaluation