Follow the Path: Hierarchy-Aware Extreme Multi-Label Classification

Cyber Alp Retreat, 23.06.2023

Natalia Ostapuk, Julien Audiffren, Ljiljana Dolamic,
Alain Mermoud, Philippe Cudré-Mauroux

Introduction: Big Picture

- Joint project with armasuisse Cyber Defence Campus (CYD).
- **CYD**: early identification of trends in the cyber area:
  - Comprehensive technology and market monitoring.
- **CYD & UniFR**:
  - **Taxonomy expansion:** build a high-quality taxonomy of technology-related concepts which can be automatically expanded [1].
  - **Semantic text tagging:** build a framework for tagging text with relevant concepts from the taxonomy.

https://www.ar.admin.ch/en/armasuisse-wissenschaft-und-technologie-w-t/cyber-defence_campus.html

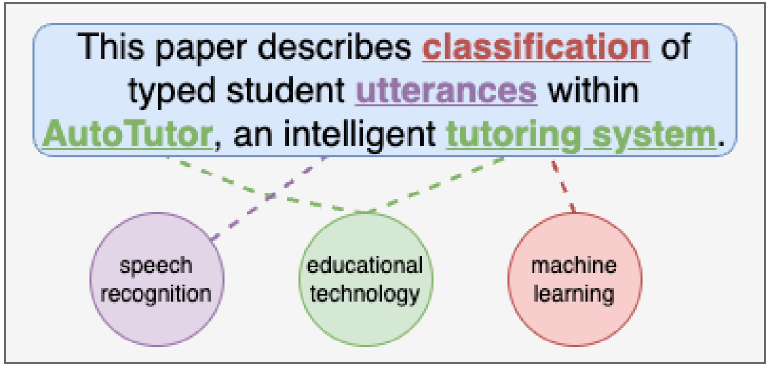
[1] Ines Arous et al. TaxoComplete: Self-Supervised Taxonomy Completion Leveraging Position-Enhanced Semantic Matching. *The WebConf, 2023.*

Introduction: Problem Statement

- **Semantic text tagging**: given a **document** in natural language and a **taxonomy** of technology related concepts, **tag the document with concepts** representing its semantic content.
- Approach: E**x**treme **M**ulti-**L**abel Text **C**lassification (**XMLC**)
  - Assigning to each document the most relevant subset of labels from an extremely large label collection.
- Specifics:
  - Scientific and technical texts.
  - Labels organized hierarchically (taxonomy).

Approach: Underlying Ideas (1/2)

Intuition 1: positive labels assigned to a document are usually represented by *specific tokens* in that document.



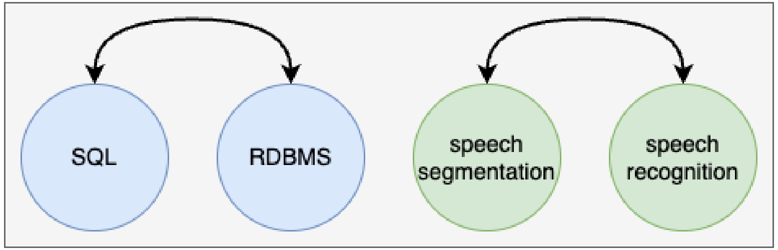
Existing XMLC approaches: document → label probabilities.

Improvement: capture *the most relevant portion of document* for each label.

Solution: attention mechanism.

Approach: Underlying Ideas (2/2)

Intuition 2: positive labels assigned to the same document are often correlated and should be treated jointly.



Existing XMLC approaches: labels are predicted *independently*.

Modeling label correlation [1]:

- Chain of binary classifiers (one per label).
- Output of each next classifier is conditioned on outputs of previous classifiers.
- Similar to the decoding process in a Seq2Seq model.

[1] Jese Read et al. TaxoComplete: Classifier Chains for Multi-Label Classification. *ECML/PKDD, 2009.*

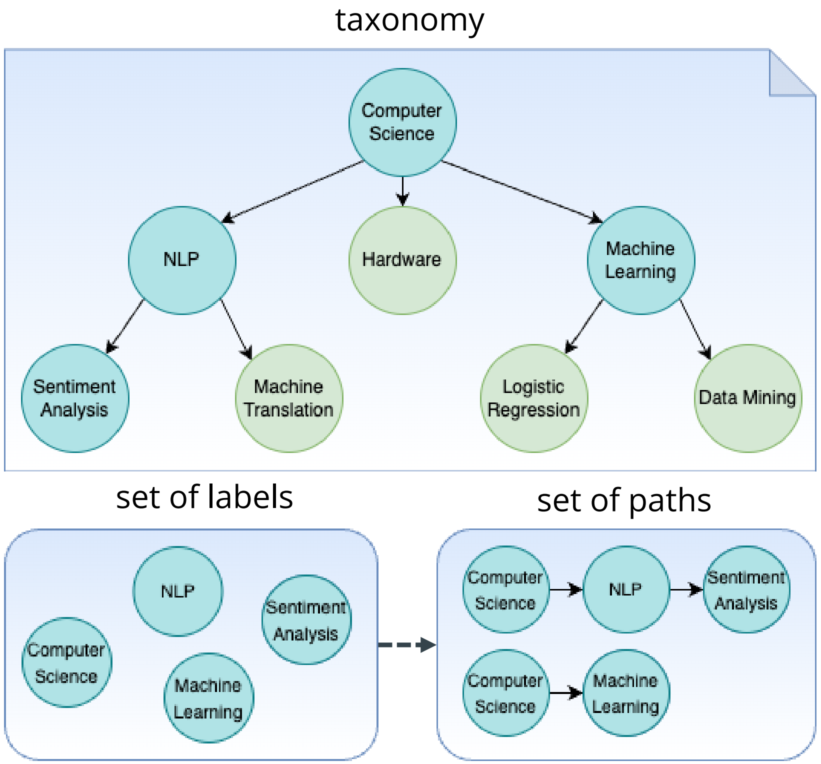
Multi-Label Classification as Seq2Seq Problem

- **Sequence-to-sequence** (Seq2Seq) learning is the task of transforming an input sequence from one domain into an output sequence from another domain.
- **Multi-Label Classification as Seq2Seq:** given an input sequence of tokens (document) generate an output sequence of labels.
- Advantages:
  - Allows to incorporate token-label cross-attention (intuition #1)
  - Labels are generated sequentially, conditioned on previously generated labels (intuition #2)
- **Issue:** labels are organized in sets and do not form a sequence.

Path Prediction

Converting labels **set** into **sequence**(s):

- Leverage label taxonomy
- Set of labels → set of paths in a taxonomy
- Each path *does* form a sequences and can be used in Seq2Seq models

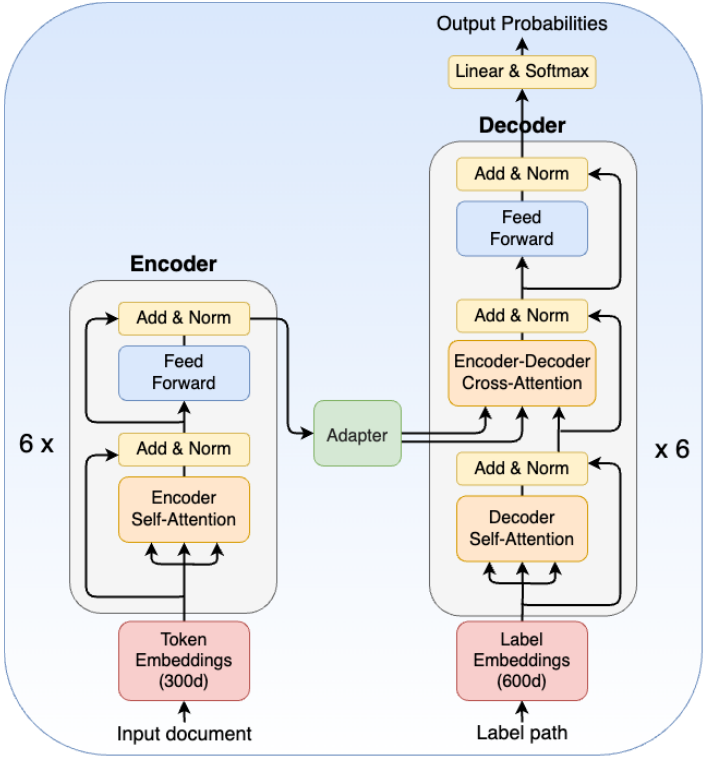


Approach: Summary

- **HECTOR** – **H**ierarchical **E**xtreme **C**lassifier for **T**ext based on transf**OR**mers.
- Label prediction → **path prediction.**
- Leverage Seq2Seq architecture – **Transformer**.
- Transformer encoder-decoder **cross-attention** highlights the most relevant tokens of the input data w.r.t. each label.
- Labels are predicted **sequentially**, from the most generic (first level of the taxonomy) to more specific.
    - Labels at top levels are easier to predict.
    - Predicted top labels then serve as an additional signal for predicting labels at deeper levels.

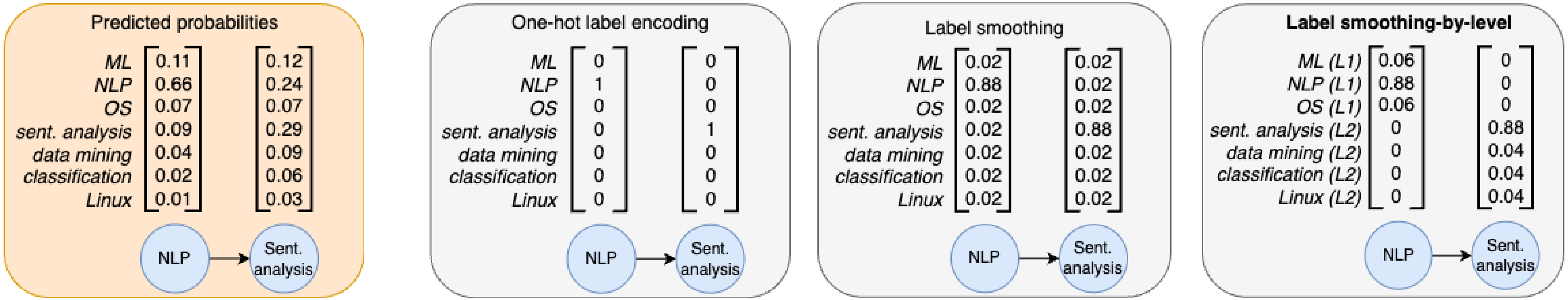
HECTOR: Architecture

- Encoder:
  - Generate contextualized representations of tokens in the input document.
- Decoder:
  - 600d label embeddings.
  - Decoder self-attention: considers previously predicted labels to generate a coherent path.
  - Encoder-decoder cross-attention: captures dependencies between input tokens and output labels.
- Prediction layer:
  - Predict label probabilities from generated label representations.



HECTOR: Loss Function

- **Kullback-Leibler divergence loss** – measures dissimilarity between two probability distributions.
- **Label smoothing** – replacing the one-hot encoding of the target labels with a smoothed distribution.
- **Label smoothing-by-level** – distribute smoothing mass throughout labels of the corresponding level of taxonomy.



HECTOR: Training

- **Multi-path problem**: labels from different (sub-)domains belong to different paths in the taxonomy => multiple relevant paths per document.
- Solution: randomly select one path at each training epoch.
  - Introduce variability and avoid overfitting.
  - Model learns to generate all possible paths with equal probability.

- **Ensemble training:**
  - Each instance of HECTOR is trained on a slightly different subset of paths => capture different aspects of data.
  - Ensemble of models improves overall quality and diversity of generated paths.

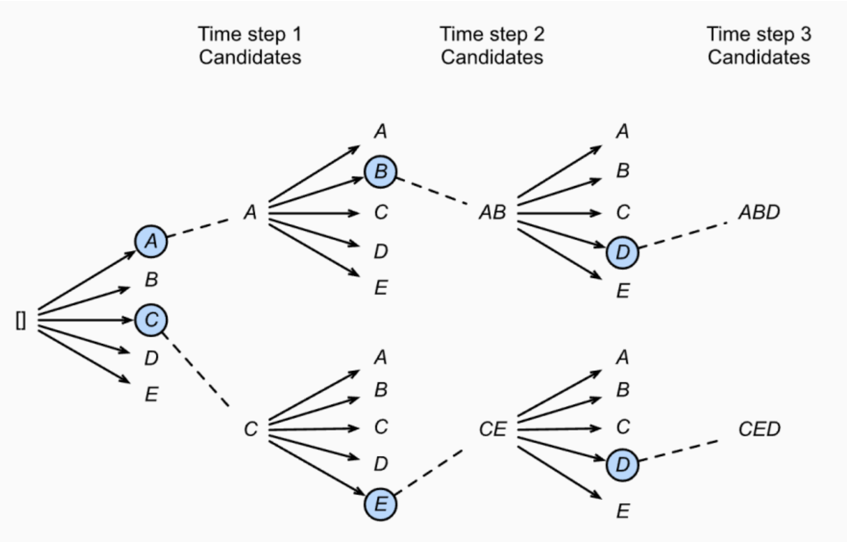
HECTOR: Path Generation

- Paths are generated by decoder.
- Sequential decoding label-by-label.
- Individual label probability:

$P(l_j, p) = \prod_{i=0}^{j} p(l_i)$ where $p = (l_0, l_1, \dots, l_{j-1})$

- Beam search algorithm:
  - Maintains a ***set*** of the most promising candidate paths.
  - Allows decoding ***multiple sequences*** simultaneously.
- <u>Final ranking:</u> sort labels from paths by their individual probabilities.

Beam search algorithm



https://d2l.ai/chapter_recurrent-modern/beam-search.html

Experiments and Results

Evaluation Tasks

- Label completion:
    - Goal: **predict missing** or incomplete labels for documents where labels are **partially provided**.
    - Motivation: subjectivity of human annotators, evolving data, privacy concerns, etc.
    - **XMLC** – specific case of label completion with **0% observed** labels.
- Label refinement:
    - Special case of label completion where only **general** labels are provided.
    - Motivation: labels representing broader categories or **higher-level concepts** are often **trivial** to predict, while assigning more **specific labels** might be **challenging**.

Datasets

- MAG-CS:
  - Dataset: abstracts of papers published at top **CS** conferences from 1990 to 2020.
  - Taxonomy: MAG label taxonomy, CS domain (descendants of *Computer Science* concept).
- PubMed:
  - Dataset: papers published in 150 top journals in **medicine** from 2010 to 2020.
  - Taxonomy: Medical Subject Headings (MeSH) hierarchically-organized thesaurus.
- EURLex:
  - Dataset: English EU **legislative documents** from the EUR-LEX portal.
  - Taxonomy: European Vocabulary (EuroVoc) multidisciplinary thesaurus.

[1] Y. Zhang, et al. "MATCH: Metadata-Aware Text Classification in a Large Hierarchy." *WWW*, 2021.
[2] Z. Shen, et al. "A Web-Scale system for scientific knowledge exploration." *ACL*, 2018.

Baselines

- **XML-CNN** – features CNN for learning an input document representation.
- **AttentionXML** – uses RNN for document representation and multi-label attention mechanism.
- **MATCH** – jointly pretrain embeddings for the metadata; leverage label hierarchy for model regularization.
- **Transformer** – vanilla Transformer encoder for document representation with a fully-connected layer for multi-label classification.

Metrics

- **NDCG@k:** measures the quality of ranking assigning higher scores to hits at top ranks; accounts for the varying number of positive labels per instance.

$$DCG@k = \sum_{l=1}^{k} \frac{y_{rank(l)}}{log(l+1)}$$

$$NDCG@k = \frac{DCG@k}{\sum_{l=1}^{min(k,||y||_0)} \frac{y_{rank(l)}}{log(l+1)}}$$

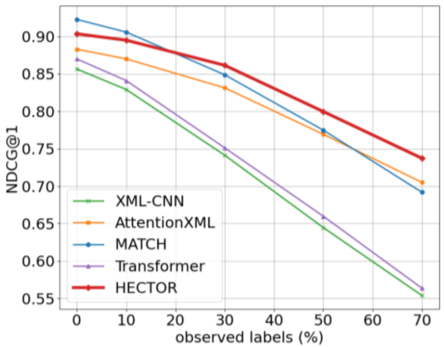
- **Precision@k:** the number of correct predictions considering only the top k elements divided by k

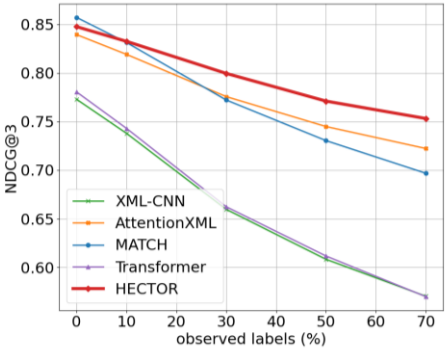
Label Completion: Experiment Design

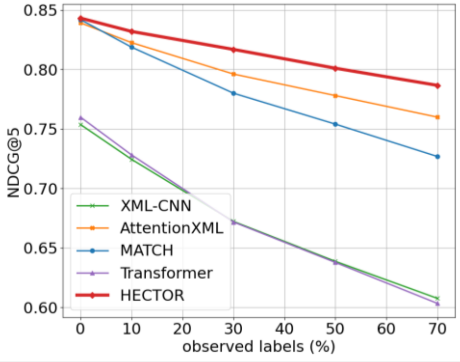
- **Assumption**: each instance in a test set contains a complete set of labels.
- Randomly drop labels:
  - $x\%$ – dropped labels
  - $(x\ -100)\%$ – observed labels
  - $x = \{30, 50, 70, 90, 100\}$
- **Task**: predict *dropped* labels.
- **Baselines**: remove observed labels from predictions.
- **HECTOR**:
  - Leverage observed labels as an additional input to the decoder (path prefixes).
  - Remove observed labels from predictions.

Label Completion: Results (1/2)

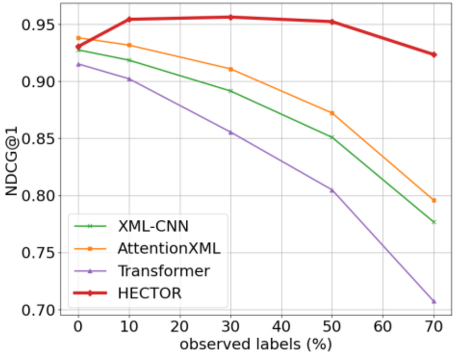
MAG-CS:

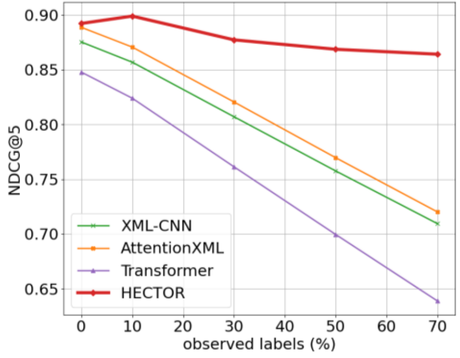


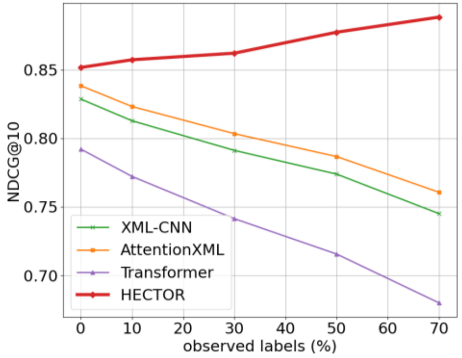




EURLex:

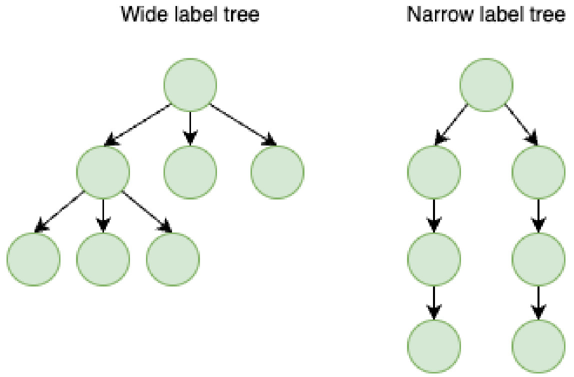






Label Completion: Results (2/2)

- Outperform existing methods at **XMLC** task on **EURLex** dataset.
- Outperform existing methods at **label completion with 10%** observed labels across **all datasets**.
- HECTOR advantage increases as more initial labels are provided.
- Performance varies across datasets:
  - Challenges:
    - Deep taxonomies (PubMed)
    - Wide label trees (MAG-CS)
  - Advantages:
    - Faster convergence (EURLex)



Label Refinement: Experiment Design

- **Assumption:** labels from the **1st** level of a taxonomy are *observed*.
- **Task:** predict labels of level 2 and deeper.
- Proportion of level 1 labels in datasets:
  - MAG-CS: 55%
  - PubMed: 13%
  - EURLex: 42%
- **Baselines**: remove all level 1 labels from predictions.
- **HECTOR**:
  - Leverage level 1 labels as path prefixes: start decoding from the 2nd position.
  - Remove all level 1 labels from predictions.

Label Refinement: Results

Algorithms	MAG-CS			PubMed			EURLex		
	nDCG@1	nDCG@3	nDCG@5	nDCG@1	nDCG@10	nDCG@20	nDCG@1	nDCG@5	nDCG@10
XML-CNN	0.6836	0.6166	0.6177	0.9187	0.8699	0.8143	0.9024	0.8168	0.8194
AttentionXML	0.8654	0.8366	0.8391	0.9288	0.8919	0.8509	0.9204	0.8329	0.8408
MATCH	0.8424	0.7782	0.7707	0.9190	0.8740	0.8166	-	-	-
Transformer	0.6670	0.5930	0.5888	0.9215	0.8718	0.8139	0.8653	0.7688	0.7732
HECTOR	0.8906	0.8515	0.8511	0.9753	0.9108	0.8955	0.9860	0.9557	0.9566
HECTOR Ens.	**0.9004**	**0.8645**	**0.8648**	**0.9777**	**0.9155**	**0.9001**	**0.9888**	**0.9597**	**0.9598**

- Outperforms all competing methods by 2.8%-6.8% (at NDCG@1).
- Reaches almost 100% at NDCG@1 on EURLex.
- **MAG-CS** and **EURLex**: label refinement results are **consistent** with label completion results.
- **PubMed**: label refinement results are **better** than label completion results.
    - 10% level 1 labels are more helpful than 10% random labels.
    - Knowing where to start, HECTOR navigates through deep taxonomies with more confidence.

Conclusion

- Introduce a **new paradigm** for **XMLC** where labels are predicted as **paths** in hierarchical label trees;
- Explore the potential of the **full Transformer** model with encoder-decoder architecture for XMLC;
- Present a new model, **HECTOR**, which is able to **capture the important portions of text** for each label and directly **leverages a label hierarchy**;
- Demonstrate the **effectiveness** of our approach for **label completion** through an extensive evaluation on three real-world XMLC datasets.

Next Steps

Zero-Shot Extreme Classification: Introduction

- Classifiers assume that the label space is fixed.
  - Typically not the case, e.g., for technology monitoring.
- **Zero-shot learning** addresses the problem of *unseen* labels which are absent during the training time.
- **Few-shot learning** addresses the problem of *tail* labels prediction (i.e. labels which are underrepresented during the training time).
- *Unseen* and *tail* labels usually provide more recent and more specific information compared to head labels.

Zero-Shot Extreme Classification: Approach

- Generative approach to zero-shot learning:
  - Synthesize the relevant labels for a given test point *starting from the given prefix*.
  - Leverage label definition and its position in the taxonomy.
- HECTOR for few-shot learning:
  - Label completion for few-shot labels (< 5 training points)

Results on MAG-CS dataset:

	MeanRank	Recall@50
XML-CNN	25.5	0.1392
AttentionXML	**19.3**	0.2533
MATCH	20.90	0.3303
HECTOR	20.59	**0.4130**

XMLC Benchmark for Scientific Document Collections

- It is still today unclear which **automatic labelling solution** should be adopted for the **TM platform**.
- Existing XMLC methods are mostly evaluated on datasets derived from Wikipedia and Amazon.
- Scientific and technical documents differ in terms of form and content.

- **Task:** develop a new **benchmark** for XMLC with a focus on **labelling scientific documents** collections.

- Implemented by an MSc student **Bhargav Solanki**

XMLC Benchmark for Scientific Document Collections

- **Datasets:**
  - MAG, PubMed
  - OpenAlex – open catalog of the world's scholarly papers.
- **Metrics:**
  - Performance: Prec@k, nDCG@k.
  - Propensity-scored performance: PSPrec@k, PSnDCG@k – place specific emphasis on performing well on rare labels.
  - Label diversity score: diversity of labels predicted by the model across all examples.
  - Efficiency: model size, training time, inference time.

Thank you!

Follow the Path: Hierarchy-Aware Extreme Multi-Label Classification

Path Completion

Observation: labels are assigned inconsistently in terms of their relative position in the taxonomy.

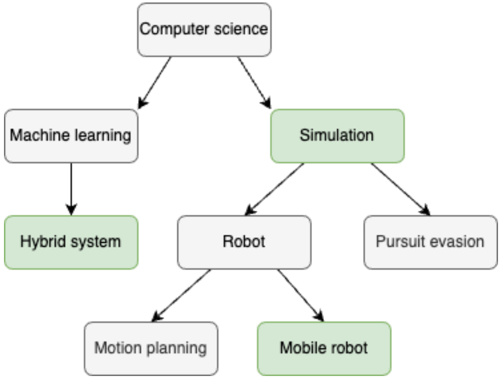
Assumption: if label ℓ_j is relevant for a document d_i, then $\boldsymbol{parent(\ell_j)}$ is also relevant for the same document.

Original data point

The design of controllers for hybrid systems in a systematic manner remains a challenging task. In this case study, we apply formal modeling to the design of communication and control strategies for a team of autonomous robots to attain specified goals in a coordinated manner.

Mobile robot | Simulation | Hybrid system

Original labels in the taxonomy



Completed labels in the taxonomy