Follow the Path: Hierarchy-Aware Extreme Multi-Label Completion for Semantic Text Tagging

Natalia Ostapuk[1], Julien Audiffren[1], Ljiljana Dolomic[2], Alain Mermoud[2] and Philippe Cudré-Mauroux[1]

[1] eXascale Infolab, University of Fribourg, Switzerland

[2] armasuisse Science and Technology, Switzerland

Cyber Alp Retreat 2024
Technology Monitoring Track

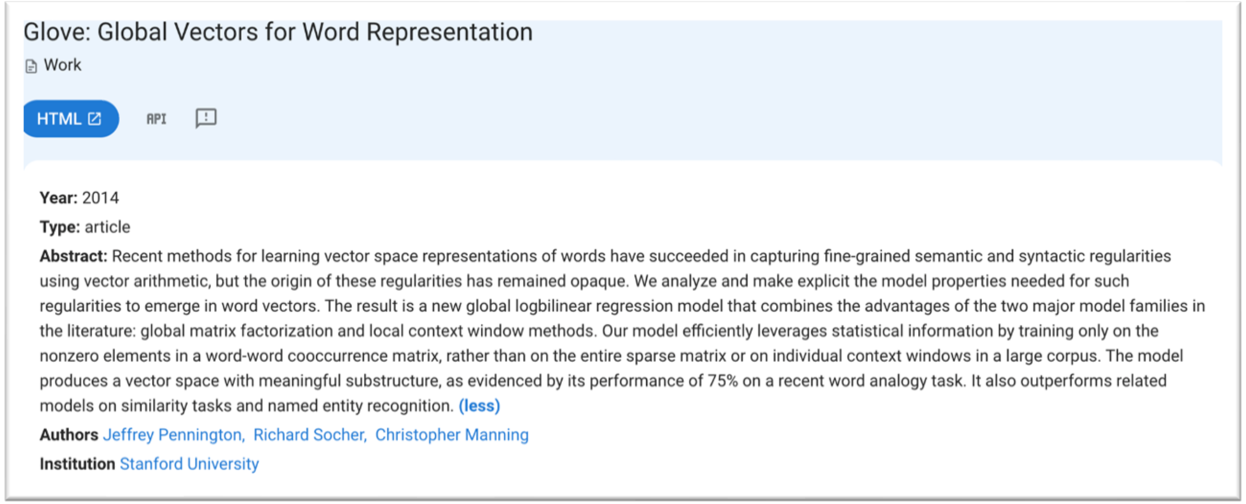
Problem Statement

- Semantic text tagging
- Extreme multi-label classification and completion

Semantic Text Tagging

The task of assigning **predefined labels** (tags) to an **entire document** or paragraph based on its content.

OpenAlex



Glove: Global Vectors for Word Representation

Work

HTML API

Year: 2014

Type: article

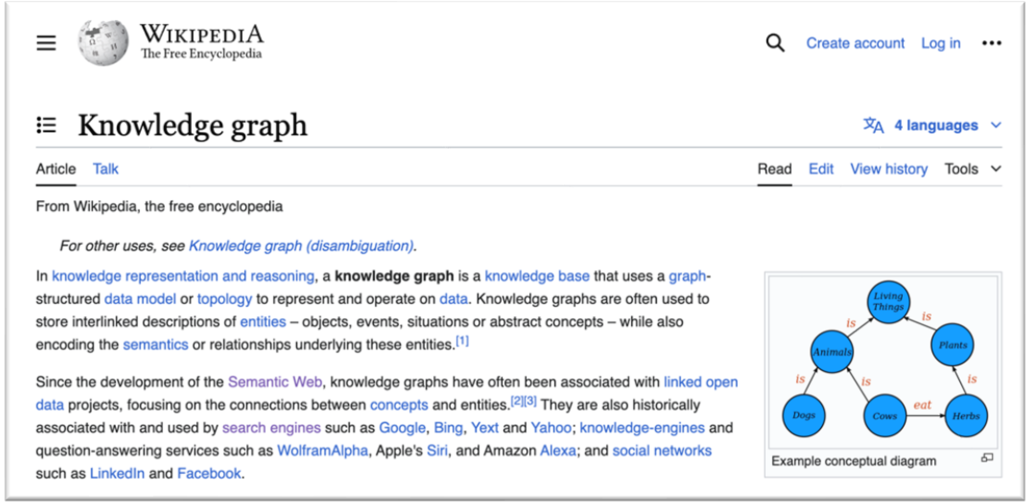
Abstract: Recent methods for learning vector space representations of words have succeeded in capturing fine-grained semantic and syntactic regularities using vector arithmetic, but the origin of these regularities has remained opaque. We analyze and make explicit the model properties needed for such regularities to emerge in word vectors. The result is a new global logbilinear regression model that combines the advantages of the two major model families in the literature: global matrix factorization and local context window methods. Our model efficiently leverages statistical information by training only on the nonzero elements in a word-word cooccurrence matrix, rather than on the entire sparse matrix or on individual context windows in a large corpus. The model produces a vector space with meaningful substructure, as evidenced by its performance of 75% on a recent word analogy task. It also outperforms related models on similarity tasks and named entity recognition. **(less)**

Authors Jeffrey Pennington, Richard Socher, Christopher Manning

Institution Stanford University

word representation | language modelling | NLP | artificial intelligence | word embeddings | ML

Wikipedia



Wikipedia
The Free Encyclopedia

Create account Log in

Knowledge graph

4 languages

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

For other uses, see Knowledge graph (disambiguation).

In knowledge representation and reasoning, a **knowledge graph** is a knowledge base that uses a graph-structured data model or topology to represent and operate on data. Knowledge graphs are often used to store interlinked descriptions of entities – objects, events, situations or abstract concepts – while also encoding the semantics or relationships underlying these entities.[1]

Since the development of the Semantic Web, knowledge graphs have often been associated with linked open data projects, focusing on the connections between concepts and entities.[2][3] They are also historically associated with and used by search engines such as Google, Bing, Yext and Yahoo; knowledge-engines and question-answering services such as WolframAlpha, Apple's Siri, and Amazon Alexa; and social networks such as LinkedIn and Facebook.

Example conceptual diagram

Categories: Knowledge graphs | Ontology (information science) | Formal semantics (natural language) | Information science

Semantic Text Tagging

- Improves computer understanding of document content
- Enhances search and discovery
- Helps data integration
- Facilitates the creation of structured representations of knowledge

Explain complexity.
Need for structure

Extreme Multi-Label Classification

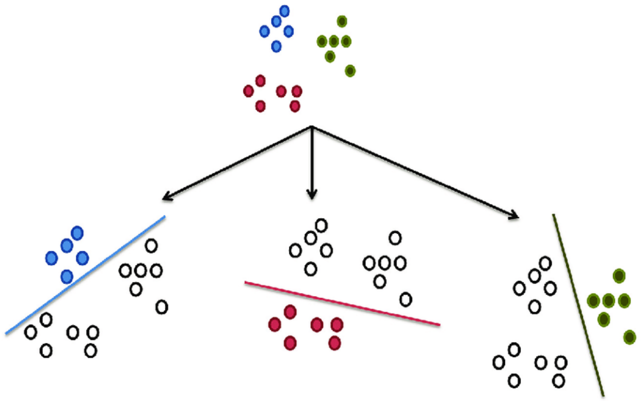
- Semantic Text Tagging is often approached as an **Extreme Multi-Label Classification** (XMLC) problem.

- Multi-label classification: assigning **multiple** labels to a single input.
- **Extreme** multi-label classification: extremely large label space (1000s – 100,000s).

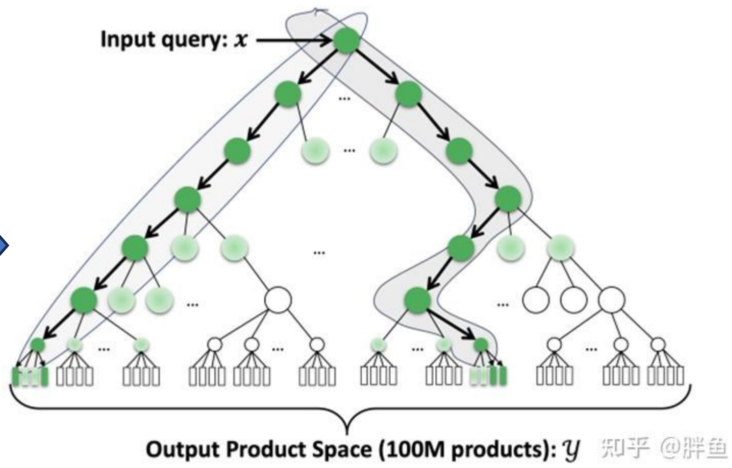
Extreme Multi-Label Classification

- Large label spaces call for large Metric/Space/Structure assumption

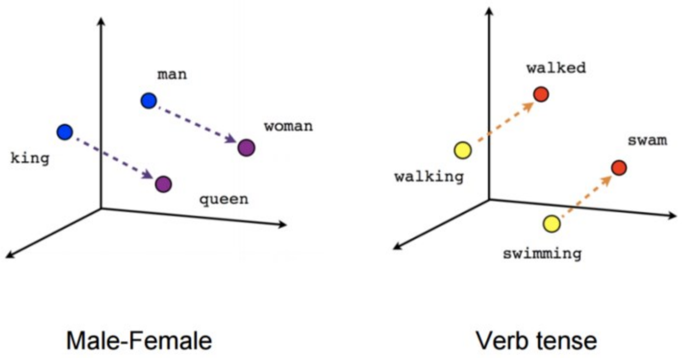
Two popular approaches



One versus All
Doesn't scale



Clustering Tree [3]



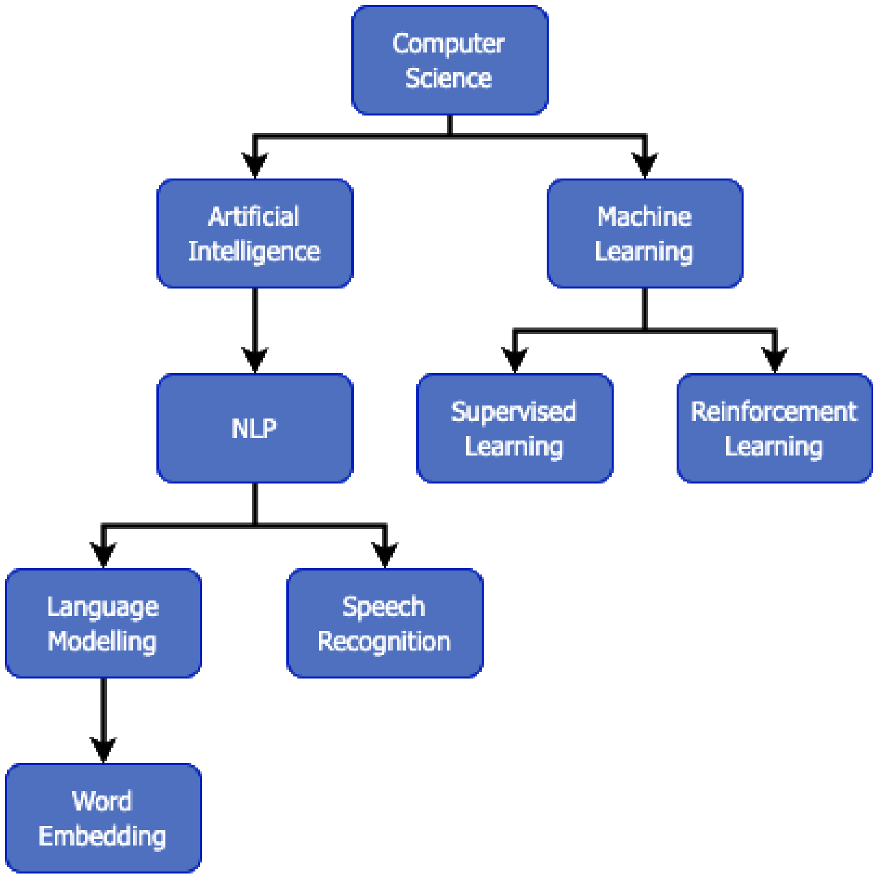
Label embedding [2]

Hierarchy of Labels

Labels are often **hierarchically** organized (taxonomy, ontology).

Provide free structural information

- Wikipedia categories
- OpenAlex – scientific concepts
- MeSH – medical subjects
- EuroVoc – EU legislation



Our Objective : Extreme Multi-Label Completion

- Sub-problem of XMLC: data instances are **partially labelled**
- Task: complete the annotation

- Hierarchically organized labels:
  - General (high-level) labels are provided
  - Specific labels are missing
  - Refinement task

Glove: Global Vectors for Word Representation

Work

HTML API

Year: 2014

Type: article

Abstract: Recent methods for learning vector space representations of words have succeeded in capturing fine-grained semantic and syntactic regularities using vector arithmetic, but the origin of these regularities has remained opaque. We analyze and make explicit the model properties needed for such regularities to emerge in word vectors. The result is a new global logbilinear regression model that combines the advantages of the two major model families in the literature: global matrix factorization and local context window methods. Our model efficiently leverages statistical information by training only on the nonzero elements in a word-word cooccurrence matrix, rather than on the entire sparse matrix or on individual context windows in a large corpus. The model produces a vector space with meaningful substructure, as evidenced by its performance of 75% on a recent word analogy task. It also outperforms related models on similarity tasks and named entity recognition. (less)

Authors Jeffrey Pennington, Richard Socher, Christopher Manning

Institution Stanford University

word representation | language modelling | NLP | artificial intelligence | word embeddings | ML

Problem Statement

Given a **document** in natural language tagged with **general labels** and a **taxonomy** of labels, predict **specific labels** for this document.

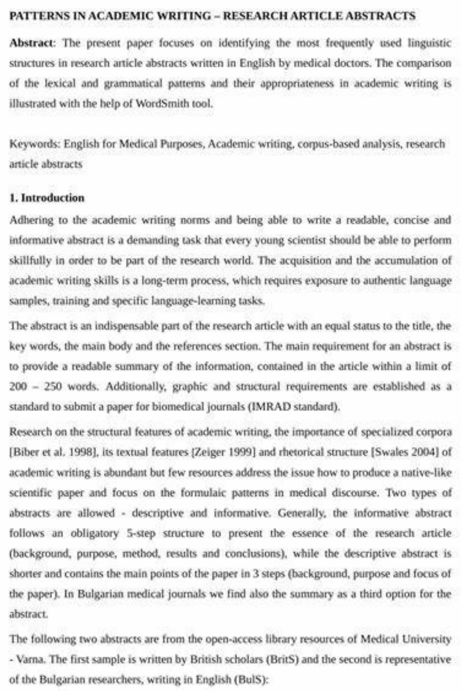
Our Approach

- Label completion as a Seq2Seq task

The main idea:
XMLC as a Seq2Seq Task

- Text has a natural sequence structure

- But what about labels ?


PATTERNS IN ACADEMIC WRITING – RESEARCH ARTICLE ABSTRACTS

Abstract: The present paper focuses on identifying the most frequently used linguistic structures in research article abstracts written in English by medical doctors. The comparison of the lexical and grammatical patterns and their appropriateness in academic writing is illustrated with the help of WordSmith tool.

Keywords: English for Medical Purposes, Academic writing, corpus-based analysis, research article abstracts

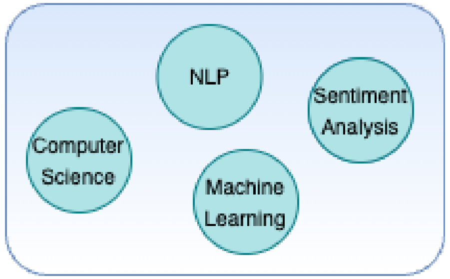
1. Introduction

Adhering to the academic writing norms and being able to write a readable, concise and informative abstract is a demanding task that every young scientist should be able to perform skillfully in order to be part of the research world. The acquisition and the accumulation of academic writing skills is a long-term process, which requires exposure to authentic language samples, training and specific language-learning tasks.

The abstract is an indispensable part of the research article with an equal status to the title, the key words, the main body and the references section. The main requirement for an abstract is to provide a readable summary of the information, contained in the article within a limit of 200 – 250 words. Additionally, graphic and structural requirements are established as a standard to submit a paper for biomedical journals (IMRAD standard).

Research on the structural features of academic writing, the importance of specialized corpora [Biber et al. 1998], its textual features [Zeiger 1999] and rhetorical structure [Swales 2004] of academic writing is abundant but few resources address the issue how to produce a native-like scientific paper and focus on the formulaic patterns in medical discourse. Two types of abstracts are allowed - descriptive and informative. Generally, the informative abstract follows an obligatory 5-step structure to present the essence of the research article (background, purpose, method, results and conclusions), while the descriptive abstract is shorter and contains the main points of the paper in 3 steps (background, purpose and focus of the paper). In Bulgarian medical journals we find also the summary as a third option for the abstract.

The following two abstracts are from the open-access library resources of Medical University - Varna. The first sample is written by British scholars (BritS) and the second is representative of the Bulgarian researchers, writing in English (BulS):



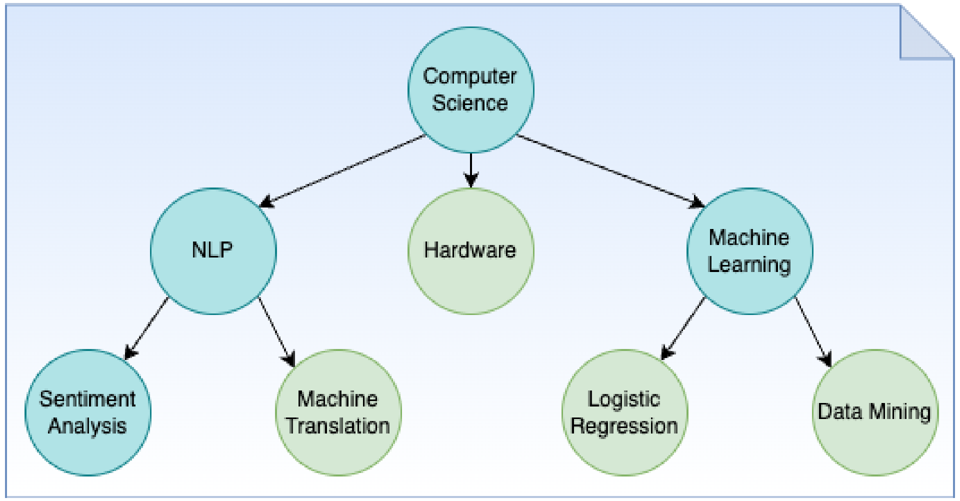
Labels

The main idea: Classification as a Seq2Seq Task

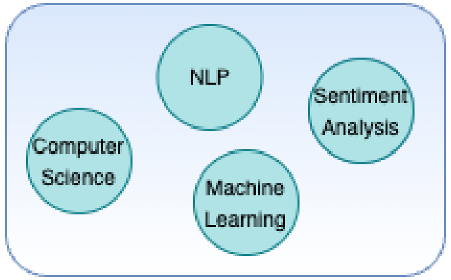
Converting label **set** into **sequence**(s):

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

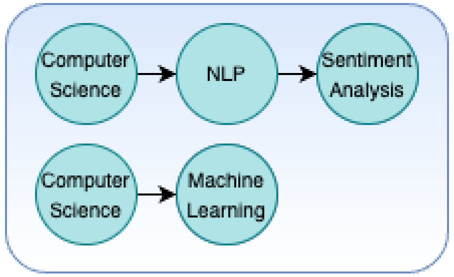
taxonomy



set of labels



set of paths



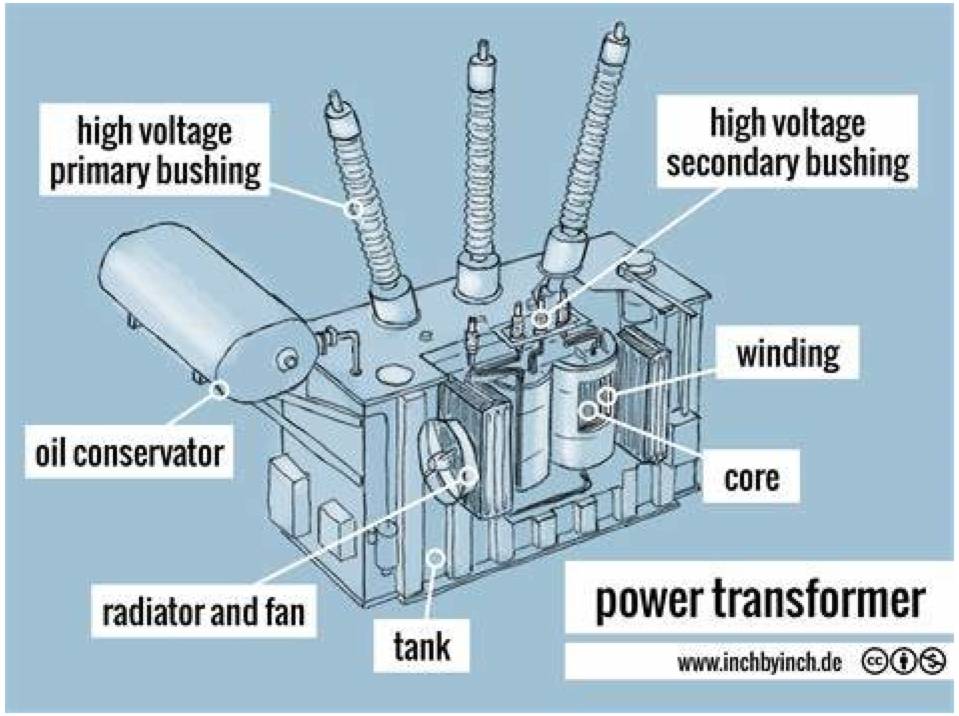
eXascale Infolab

UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

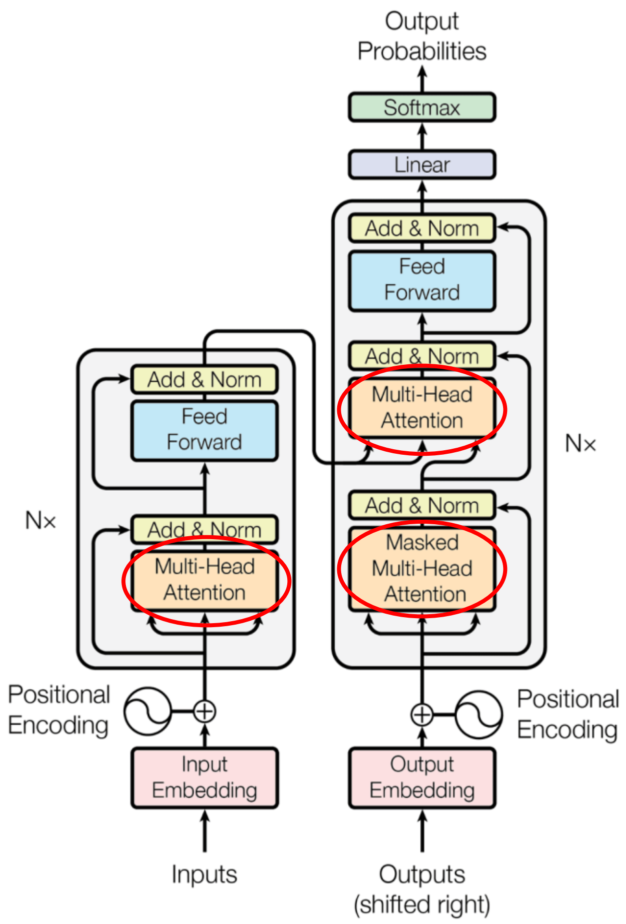
What is good at Seq2Seq ?

💡

Transformers!



Transformers for XMLC



Consider labels as a target sequence.

☑ Encoder-decoder cross-attention: label representation w.r.t. tokens from input document.

☑ Decoder self-attention: contextualized label representation.

🎁 Encoder self-attention: contextualized word representation (useful for GloVe embeddings).

Problem: labels are organized in sets and do not form a **sequence**.

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.

Experiments and Results

- Experiment design
- Datasets and baselines
- Experimental results

Experiment Design

- Label refinement task:
  - **Initial state:** Each document is tagged with a **partial set** of labels, corresponding to **top level(s)** of taxonomy.
  - **Goal**: predict more **specific labels.**
  - **Experiment parameter:** **level L**, from which the refinement begins.
    - Document is tagged with labels of from level 1 to L - 1
  - XMLC is a special case of label refinement with $L == 1$.

Datasets

- MAG-CS [3]:
  - Dataset: abstracts of papers published at top **CS** conferences from 1990 to 2020.
  - Taxonomy: MAG label taxonomy, CS domain (descendants of *Computer Science*).
- PubMed [3]:
  - Dataset: papers published in 150 top journals in **medicine** from 2010 to 2020.
  - Taxonomy: Medical Subject Headings (MeSH) hierarchically-organized thesaurus.
- EURLex [6]:
  - Dataset: English EU **legislative documents** from the EUR-LEX portal.
  - Taxonomy: European Vocabulary (EuroVoc) multidisciplinary thesaurus.

Baselines

- Extreme multi-label classification models:
  - XML-CNN [1]
  - AttentionXML [2]
  - MATCH [3]
  - XR-Transformer [4]
- Multi-label completion models:
  - REASSIGN [5]

Results: Ranking Metrics

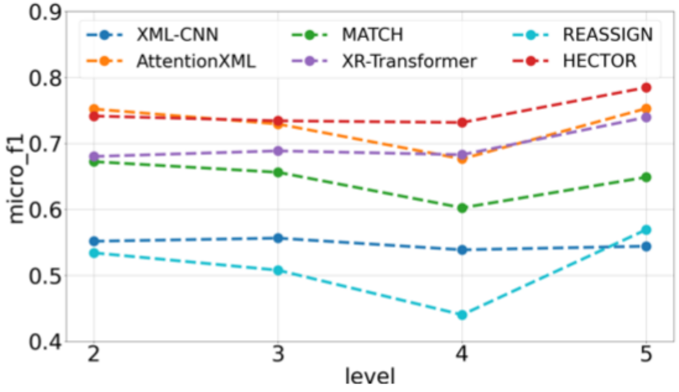
L	Algorithms	MAG-CS				PubMed				EURLex			
		P@1	P@3	N@3	N@5	P@1	P@3	N@3	N@5	P@1	P@3	N@3	N@5
2	XML-CNN	0.7002	0.4516	0.6366	0.6390	0.9190	0.8942	0.9026	0.8902	0.8998	0.8136	0.8471	0.8147
	AttentionXML	0.8665	0.5884	0.8381	0.8406	0.9288	0.9103	0.9175	0.9082	0.9205	0.8344	0.8676	0.8334
	MATCH	0.8434	0.5363	0.7795	0.7721	0.9190	0.8967	0.9047	0.8937	-	-	-	-
	XR-Transformer	0.8027	0.5437	0.7677	0.7717	0.9180	0.9041	0.9104	0.9029	0.9276	0.8587	0.8890	0.8568
	REASSIGN	0.6680	0.4224	0.5942	0.5901	0.9196	0.8554	0.8713	0.8417	0.8655	0.773	0.8061	0.7691
	HECTOR	**0.8917**	**0.5931**	**0.8530**	**0.8527**	**0.9753**	**0.9436**	**0.9554**	**0.9392**	**0.9861**	**0.9419**	**0.9691**	**0.9563**
3	XML-CNN	0.6747	0.4121	0.6681	0.6913	0.8993	0.8638	0.8775	0.8681	0.8028	0.5038	0.7942	0.8146
	AttentionXML	0.8346	0.4973	0.8290	0.8448	0.9177	0.887	0.9006	0.8925	0.8220	0.5158	0.8111	0.8345
	MATCH	0.7818	0.4496	0.7583	0.7725	0.9025	0.8691	0.8827	0.8737	-	-	-	-
	XR-Transformer	0.7906	0.4770	0.7879	0.8015	0.9093	0.8827	0.8960	0.8892	0.8441	0.5211	0.8239	0.8343
	REASSIGN	0.6019	0.3636	0.5836	0.6025	0.8916	0.8301	0.8484	0.8238	0.7598	0.4791	0.7522	0.7735
	HECTOR	**0.8818**	**0.5141**	**0.8745**	**0.8885**	**0.9754**	**0.9363**	**0.9589**	**0.9468**	**0.9579**	**0.6034**	**0.9506**	**0.9595**
4	XML-CNN	0.6662	0.3777	0.7358	0.7724	0.8743	0.8547	0.8650	0.8571	0.8115	0.3690	0.8655	0.8794
	AttentionXML	0.8113	0.4257	0.8581	0.8788	0.9021	0.8816	0.8944	0.8884	0.8251	0.3775	0.8836	0.8957
	MATCH	0.7330	0.3843	0.7789	0.8071	0.8820	0.8627	0.8747	0.8678	-	-	-	-
	XR-Transformer	0.7775	0.4083	0.8197	0.8364	0.8980	0.8765	0.8907	0.8846	0.8163	0.3448	0.8289	0.8360
	REASSIGN	0.5416	0.3174	0.6015	0.6478	0.8716	0.8469	0.8584	0.8476	0.7636	0.3613	0.8359	0.8518
	HECTOR	**0.8494**	**0.4390**	**0.8961**	**0.9140**	**0.9711**	**0.9294**	**0.9601**	**0.9523**	**0.9177**	**0.3991**	**0.9542**	**0.9583**
5	XML-CNN	0.7815	0.3376	0.8581	0.8736	0.8926	0.8742	0.8871	0.8742	0.9640	0.3393	0.9739	0.9774
	AttentionXML	0.8612	0.3492	0.9101	0.9209	0.9203	0.8975	0.9150	0.9072	0.9640	**0.3483**	0.9841	0.9841
	MATCH	0.7802	0.3256	0.8368	0.8585	0.9026	0.8788	0.8962	0.8877	-	-	-	-
	XR-Transformer	0.8213	0.3243	0.8551	0.8664	0.9139	0.8891	0.9077	0.8997	0.9189	0.3273	0.9346	0.9480
	REASSIGN	0.7121	0.3205	0.8022	0.8283	0.8912	0.8723	0.8857	0.8759	0.9279	0.3393	0.9611	0.9659
	HECTOR	**0.8946**	**0.3526**	**0.9292**	**0.9370**	**0.9788**	**0.9359**	**0.9711**	**0.9610**	**0.9989**	**0.3483**	**0.9978**	**0.9978**

Starting with L **== 2,** HECTOR outperforms all competing methods by a large margin.

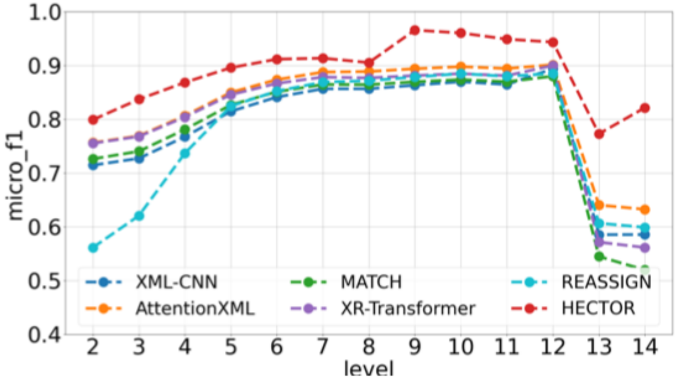
The advantage increases as more initial labels are provided.

Performance varies across datasets (challenge: wide label trees).

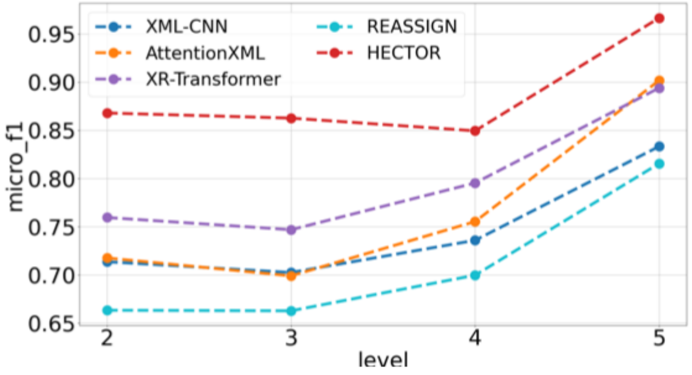
Results: Classification Metrics



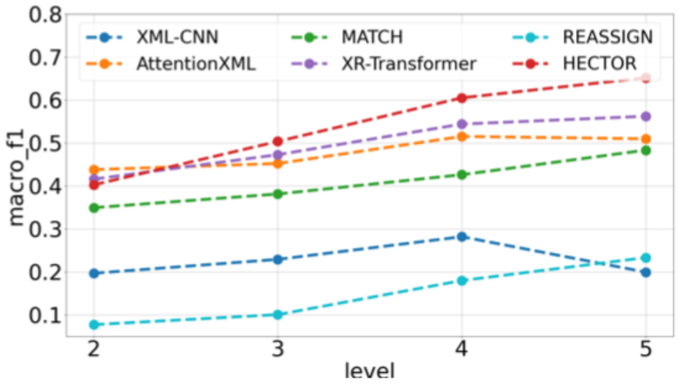
(a) MAG-CS - Micro F1



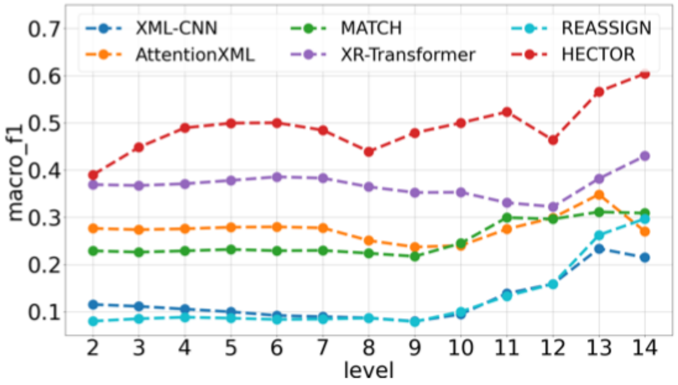
(b) PubMed - Micro F1



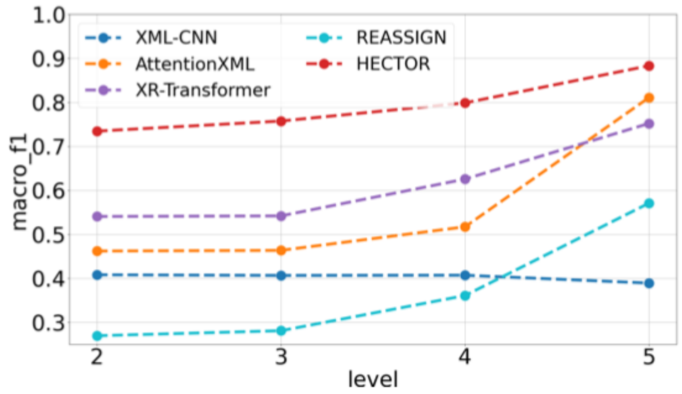
(c) EURLex - Micro F1



(d) MAG-CS - Macro F1



(e) PubMed - Macro F1



(f) EURLex - Macro F1

Conclusion

Conclusion

- Introduce a **new paradigm** for **XMLC** where labels are predicted as **paths** in a hierarchical label tree;
- 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.

References

1. J. Liu et al., *Deep learning for extreme multi-label text classification,* in SIGIR, 2017.
2. Kush Bhatia, Himanshu Jain, Purushottam Kar, Manik Varma, and Prateek Jain. 2015. Sparse Local Embeddings for Extreme Multi-label Classification. In Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada. 730–738.
3. R. You et al., *Attentionxml: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification*, in NeurIPS, 2019.
4. Y. Zhang, et al., *MATCH: Metadata-Aware Text Classification in a Large Hierarchy*, in *WWW*, 2021.
5. J. Zhang et al., *Fast Multi-Resolution Transformer Fine-tuning for Extreme Multi-label Text Classification*, in NeurIPS, 2021.
6. M. Romero et al., *Leveraging class hierarchy for detecting missing annotations on hierarchical multi-label classification,* in Comput. Biol. Medicine, 2023.
7. I. Chalkidis et al., *Large-Scale Multi-Label Text Classification on EU Legislation*. In ACL, 2019.

eXascale Infolab

UNI FR UNIVERSITÉ DE FRIBOURG UNIVERSITÄT FREIBURG

Thank you!

Follow the Path: Hierarchy-Aware Extreme Multi-Label Completion for Semantic Text Tagging

https://exascale.info