UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

Extreme Multi-Label Text Classification for Technology Monitoring

Julien Audiffren
Christophe Broilet
Philippe Cudré-Mauroux

eXascale Infolab, University of Fribourg
Switzerland

eXascale Infolab

Cyber Alp Retreat
June 2025

Text Classification for Tech. Monitoring

Capture which technologies (from a given set of relevant technologies, often organized as a taxonomy) entities are working on based on their websites (/ research papers / other documents…)

Previous contributions include

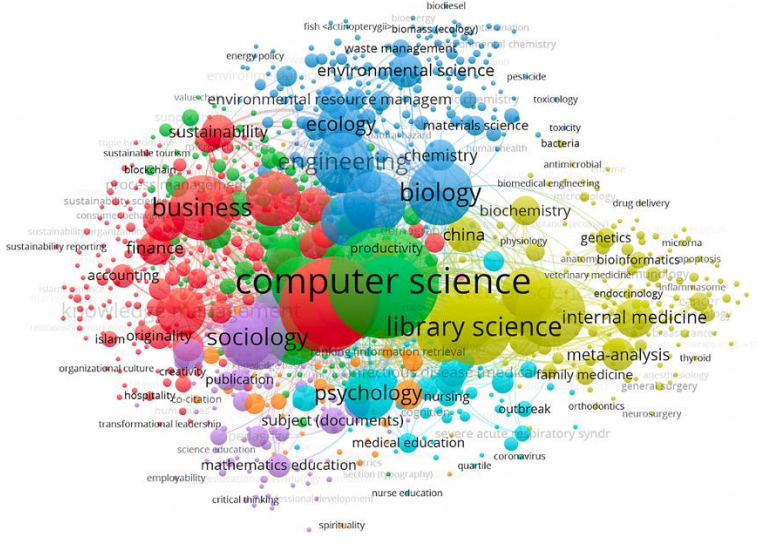
- Leveraging human rationales for explainable text classification [AAA2021]
- Self-supervised taxonomy completion for new technologies [WWW2023]
- Multi-label completion for semantic text tagging [WWW2024]

Extreme Multi-Label Classification (XMLC)

- Recent focus on very large taxonomies of labels
  - E.g., from OpenAlex

OpenAlex

- Traditional classification algorithms are not effective in such setups
  - Tens to hundreds of thousands of labels
  - Many labels only have few training instances



=> XMLC techniques

Today's Contributions

1. TAMLEC: XMLC tagging with taxonomy-aware parallel learning

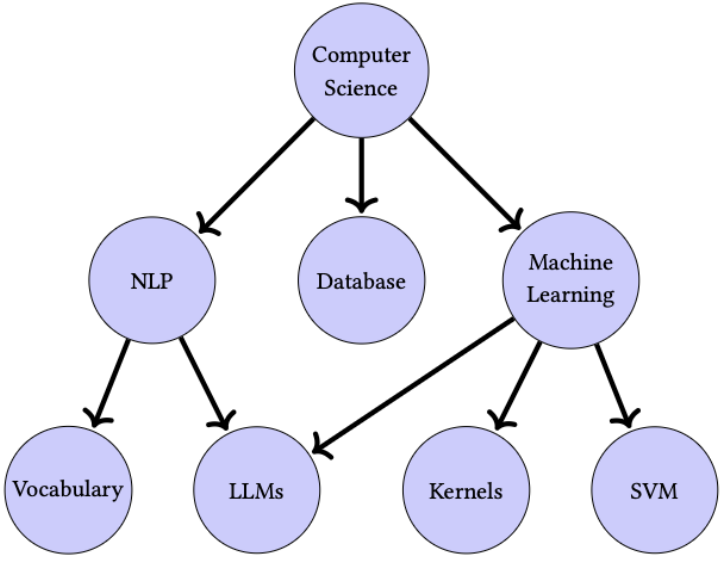
2. OAXMLC: a dual-taxonomy dataset for XMLC

3. Demo on armasuisse's labels of interest

TAMLEC (1/2)

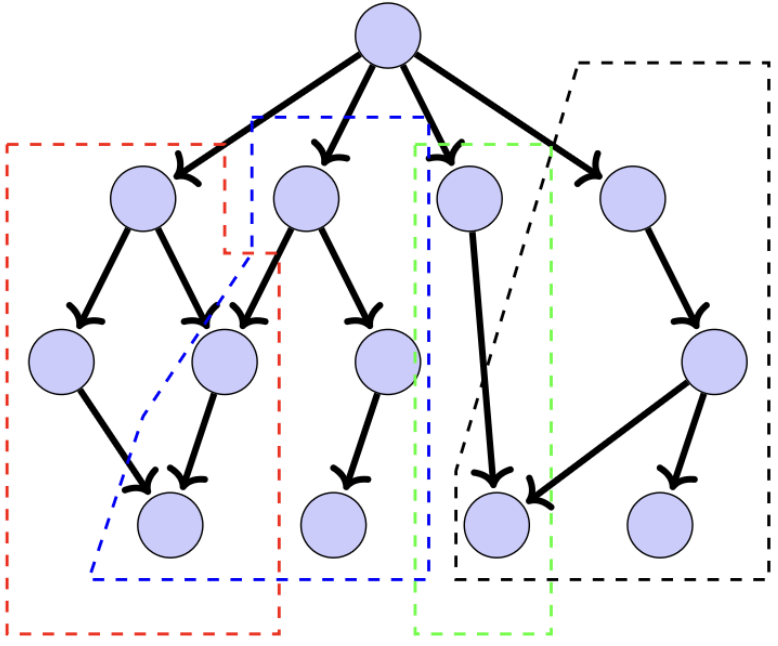
- Leveraging the taxonomic structure is often key to improving XMLC performance
  - Subsumption (superclass / subclass) relationships between labels in the taxonomy
  - *Paths* of labels (esp. useful for label completion)

- However, real taxonomies are often more complex than trees
  - Children can have more than one parent!



TAMLEC (2/2)

- TAMLEC considers weak semi-lattice taxonomic structures (which are more generic than trees)
  - It introduces the Taxonomy-Aware Tasks (TATs) decomposition, where the taxonomy is decomposed into subtasks that satisfy coherence and separability criteria
  - It then predicts paths of labels using a full Transformer architecture leveraging a TAT-dependent loss function
  - For each path and TAT, it generates multiple path extensions using beam search which are ultimately combined



Results

- Outperforms the state of the art on three standard XMLC datasets
- Superior results on label completion and in few-shot settings also

	Method	Precision				NDCG				F1			
		@1	@2	@3	@4	@2	@3	@4	@5	@1	@2	@3	@4
MAG-CS	MATCH	0.79	0.77	0.75	0.74	0.79	0.78	0.79	0.78	0.65	0.65	0.66	0.68
	X-CNN	0.52	0.56	0.56	0.57	0.58	0.61	0.63	0.64	0.41	0.46	0.49	0.52
	Att.XML	0.81	0.79	**0.78**	**0.78**	0.81	0.81	0.81	0.81	0.67	0.67	0.69	**0.71**
	HECTOR	0.83	0.76	0.74	0.72	0.79	0.77	0.77	0.76	0.69	0.65	0.65	0.66
	TAMLEC	**0.87**	**0.80**	0.78	0.77	**0.83**	**0.81**	**0.81**	**0.81**	**0.73**	**0.69**	**0.69**	0.70
PUBMed	MATCH	0.79	0.75	0.76	0.78	0.76	0.77	0.79	0.81	0.28	0.42	0.49	0.52
	X-CNN	0.75	0.70	0.70	0.73	0.71	0.72	0.75	0.78	0.26	0.38	0.44	0.48
	Att.XML	0.79	0.76	0.76	0.78	0.76	0.77	0.80	0.82	0.28	0.43	0.49	0.52
	HECTOR	0.84	0.80	0.79	0.80	0.81	0.81	0.82	0.83	0.31	0.47	0.52	0.54
	TAMLEC	**0.87**	**0.84**	**0.84**	**0.86**	**0.84**	**0.85**	**0.87**	**0.89**	**0.32**	**0.49**	**0.55**	**0.58**
EURLex	MATCH	0.85	0.89	0.85	0.84	0.89	0.86	0.85	0.79	0.75	0.84	0.81	0.81
	X-CNN	0.84	0.88	0.85	0.84	0.89	0.86	0.85	0.80	0.74	0.84	0.81	0.81
	Att.XML	0.87	0.89	0.86	0.84	0.90	0.87	0.86	0.81	0.76	0.85	0.82	0.82
	HECTOR	0.89	0.88	0.85	0.80	0.89	0.86	0.82	0.75	0.78	0.84	0.81	0.78
	TAMLEC	**0.94**	**0.95**	**0.93**	**0.90**	**0.96**	**0.94**	**0.91**	**0.85**	**0.83**	**0.91**	**0.89**	**0.87**

OAXMLC:
A Dual-Taxonomy Dataset for XMLC

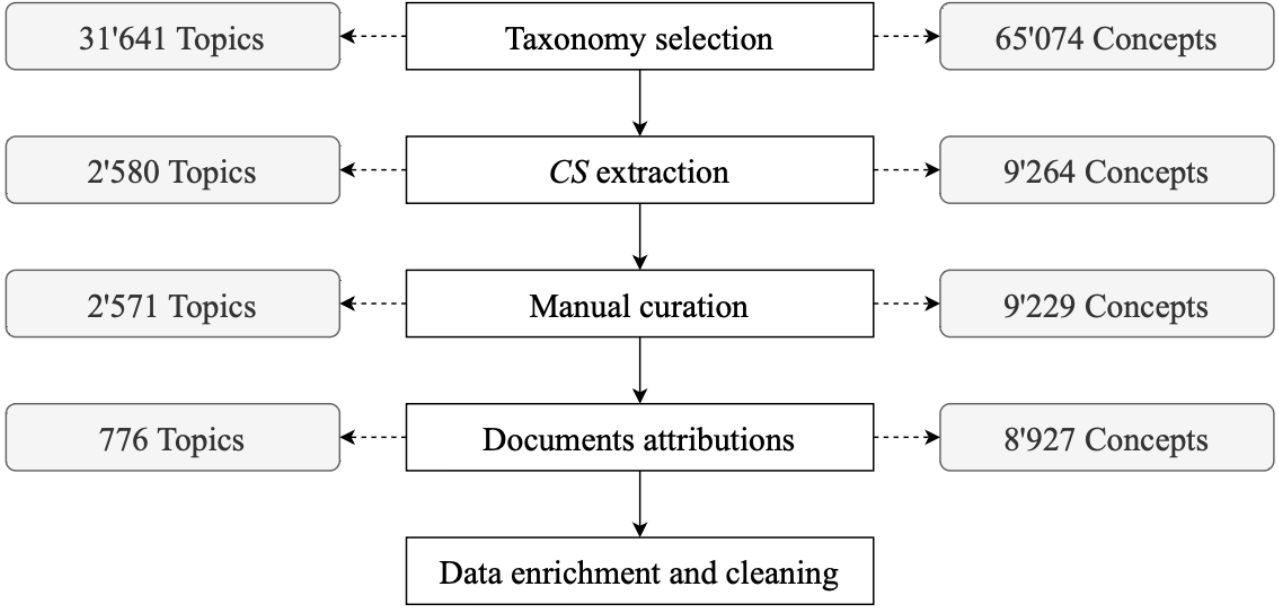
- Standard XMLC datasets
  - Target very different domains
  - Are relatively small
  - Only come with one taxonomic structure

- OAXMLC is a new XMLC dataset solving all these issues
  - Larger than previous datasets (in all dimensions)
  - Comes with two taxonomies (great to study the influence of taxonomy on training)

OAXMLC Design

- Based on OpenAlex
  - 1-2 orders of magnitude more documents
- Built using a new generic methodology
  - Involving both LLMs and human input for data annotation and cleaning
- Comes with two taxonomies
- Openly available on Zenodo
  - Taxonomies are serialized both in SKOS (Turtle) and JSON

OAXMLC Construction



OAXMLC Properties

Name	Description
`identifier`	A label to uniquely identify a node entry in the taxonomy. For the Concepts taxonomy, this corresponds to the OpenAlex concept ID, e.g. `C136764020`. For the Topics taxonomy, this corresponds to (i) the OpenAlex ID for the subfields and the topics, and (ii) to an arbitrary ID for the keywords.
`skos:prefLabel`	The textual title or name of the taxonomy entry.
`skos:definition`	The textual definition of the taxonomy entry. If no definition is available in OpenAlex, it was generated it with a LLM.
`skos:narrower`	A label more specific, according to the hierarchical taxonomy.
`oaxmlc:is_def_ai_generated`	A flag that tells if the definition was generated with a LLM or not.
`oaxmlc:is_def_from_openalex`	A flag that tells if the definition is coming from OpenAlex or not.

`oaxmlc:wikipedia`
`oaxmlc:wikidata`
`oaxmlc:OpenAlex`
	A link to the relevant URLs for this entry. It can include OpenAlex, Wikidata, and Wikipedia.

OAXMLC vs SOTA Datasests

	Documents	Labels	LPD		DPL	
	N	N	Avg	Med	Avg	Med
OAXMLC	3'725'870	-	-	-	-	-
topics	-	775	3.6	3	17132.1	1538
concepts	-	8'926	9.8	9	4100.1	362
MAG-CS	143'928	2'641	4.4	4	239.6	13
EURLex	51'000	4'492	10.4	10	118.4	11
PubMed	139'932	5'911	18.5	18	437.6	15

OAXMLC Use

- Used to get new insight on XMLC algorithms
  - Results on larger datasets
  - Dependencies from taxonomic structures

	Method	$P@1$	$P@5$	$P@10$	$N@5$	$N@10$	μP	μRec	$\mu F1$	MP	$MRec$	$MF1$
Topics	ATT.XML	0.751	**0.750**	**0.772**	**0.795**	**0.799**	0.667	0.621	0.643	0.594	0.490	0.537
	HECTOR	**0.765**	0.658	0.708	0.710	0.745	0.446	0.577	0.503	0.384	**0.523**	0.443
	MATCH	0.763	0.745	0.727	0.791	0.756	**0.668**	**0.639**	**0.653**	**0.615**	0.496	**0.549**
	XML-CNN	0.692	0.652	0.681	0.709	0.716	0.604	0.551	0.577	0.482	0.361	0.413
	FASTXML	0.390	0.366	0.374	0.405	0.402	0.433	0.251	0.317	0.323	0.080	0.128
	CASC.XML	0.750	0.724	0.759	0.775	0.791	0.648	0.620	0.634	0.553	0.451	0.497
	LIGHTXML	0.748	0.723	0.761	0.774	0.791	0.651	0.614	0.632	0.569	0.452	0.504
	PARABEL	0.439	0.454	0.483	0.506	0.520	0.360	0.264	0.304	0.277	0.190	0.225
Concepts	ATT.XML	**0.889**	**0.850**	**0.830**	**0.881**	**0.873**	0.711	**0.695**	**0.703**	**0.542**	0.394	0.456
	HECTOR	0.814	0.670	0.645	0.719	0.705	0.515	0.495	0.505	0.285	0.229	0.254
	MATCH	0.886	0.847	**0.830**	0.878	**0.873**	0.713	0.691	0.702	0.524	**0.415**	**0.463**
	XML-CNN	0.829	0.737	0.720	0.785	0.783	0.617	0.552	0.582	0.370	0.234	0.287
	FASTXML	0.513	0.490	0.540	0.531	0.599	0.562	0.242	0.338	0.190	0.032	0.055
	CASC.XML	0.863	0.789	0.757	0.831	0.816	**0.777**	0.504	0.612	0.408	0.181	0.251
	LIGHTXML	0.872	0.820	0.796	0.855	0.846	0.679	0.644	0.661	0.454	0.324	0.378
	PARABEL	0.554	0.508	0.529	0.553	0.594	0.444	0.311	0.366	0.272	0.088	0.133

DEMO

Conclusions

- XMLC is a very promising paradigm for Technology Monitoring
  - Fine-grained text classification following your needs!
- We recently made two important contributions to the XMLC field
  - A new XMLC algorithm, TAMLEC, which outperforms the state of the art in the field
  - A new dataset, OAXMLC, which has unique properties to benchmark XMLC algorithms

Thanks a lot for your attention!