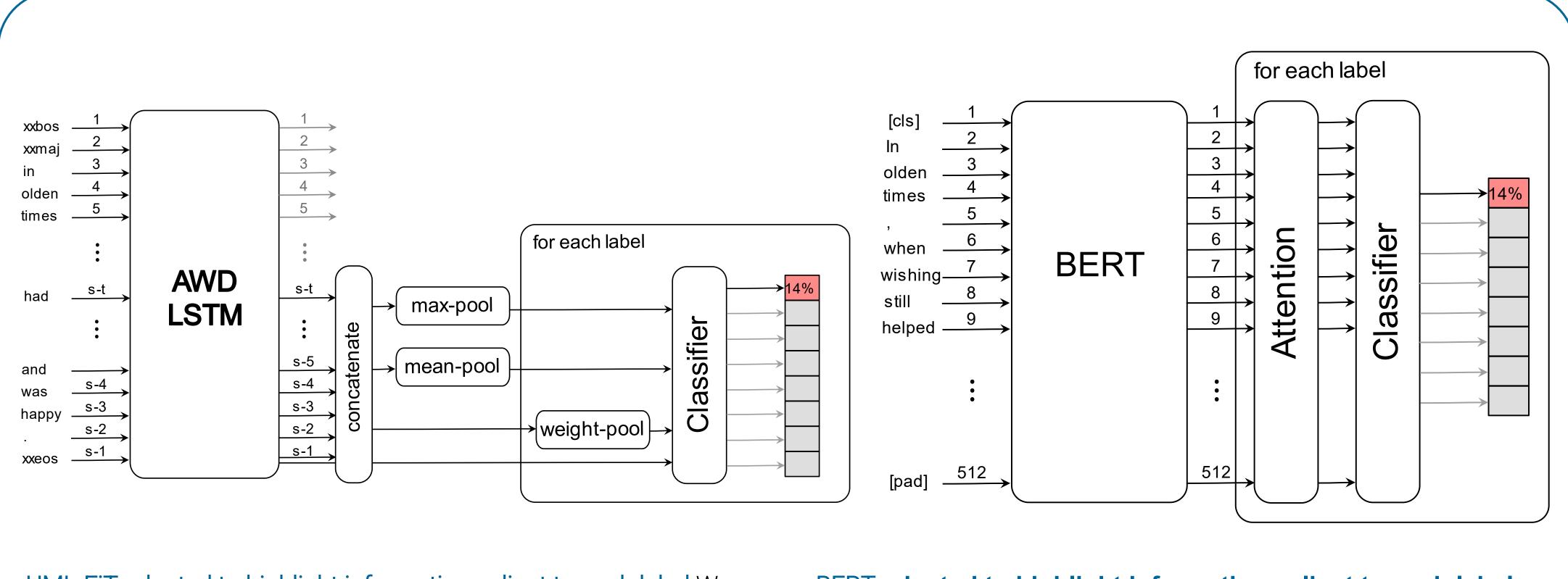


Enabling Science via Analytical Informatics

Introduction

The ECOTOXicology Knowledgebase (ECOTOX) is a comprehensive, publicly available resource providing single chemical environmental toxicity data on aquatic life, terrestrial plants and wildlife. The database is updated quarterly, and to identify relevant references and extract pertinent data, the ECOTOX data curation pipeline employs a methodical process similar to the initial stages of systematic review. This labor-intensive workflow requires curators to regularly evaluate tens of thousands of candidate references, the majority of which are then rejected as not relevant. After the careful review of hundreds of thousands of potentially relevant articles, the ECOTOX database currently (as of March 2022) contains data for 12,485 chemicals and 13,709 species manually extracted from 53,020 references. The availability of this extensive dataset of historical screening decisions provided us with the opportunity to develop high performance, stateof-the-art neural network classifiers to partially automate title and abstract screening and to categorize (e.g., human health, fate, chemical methods) rejected references. https://cfpub.epa.gov/ecotox/

Note: The views expressed in this poster are those of the authors and do not necessarily reflect the views or policies of the U.S. EPA.



UML-FiT adapted to highlight information salient to each label We modify the classification layer of UML-FiT by adding the weightpooled hidden state vector, using weights derived from word-level attention (Yang 2016, doi: 10.18653/v1/n16-1174). The attention weights are trained separately for each label.

Improving the Efficiency of Literature Identification for the ECOTOXicology Knowledgebase Using Deep Learning

BE Howard¹, CR Norman¹, A Tandon¹, R Shah¹ J Olker², D Hoff ² | 1. Sciome LLC, Research Triangle Park, NC | 2. U.S. EPA, Duluth MN

Material and Methods

We experimented on a subset of 88,900 articles spanning nearly 100 chemicals from the ECOTOX database. Out of these, 65,553 were excluded after manual screening, and annotated with a reason for exclusion.

References were cleaned prior to processing:

- Encoding problems were corrected and html entities were normalized
- Boilerplate text such as 'Abstract:', 'All rights reserved', et c. were stripped using regexes
- Abstracts were parsed to strip keywords from the abstract body
- Abstracts were parsed to strip copyright statements from the abstract body
- Keywords, journal, and publisher information were each given their own dedicated field

We trained modified versions of ULM-FiT and BERT-large to identify the 22 most frequent exclusion criteria in ECOTOX, as well the other criteria in one label (OTHER).

The final model used in ECOTOX is a hybrid meta-model which delegates decision to either UML-FiT (Howard 2018, arXiv:1801.06146) or BERT (Devlin 2019, doi 10.18653/v1/N19-1423) depending on which performs best on each label. References without abstracts are processed by a dedicated titles-only classifier.

BERT adapted to highlight information salient to each label We modify the classification layer of BERT by replacing the preclassification layer with a scaled dot-product attention head (Vaswani 2017, arXiv:1706.03762). The attention heads are trained separately for each label.

Results Screene Screer Screene Screene Screener 25

Table 1: Final performance, cross validated over 5 folds. P denotes precision, R denotes recall (sensitivity), F denotes the F₁ measure.

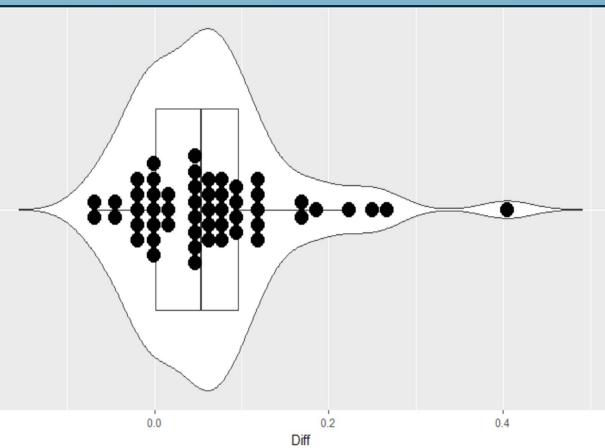
	Р	R	F
HUMAN HEALTH	67.09%	60.06%	55.30%
CHEM METHODS	77.61%	73.16%	74.51%
FATE	55.01%	70.04%	58.29%
BACTERIA	65.30%	66.41%	64.71%
REVIEW	71.78%	66.98%	69.07%
SURVEY	61.71%	65.99%	63.50%
MIXTURE	51.22%	60.43%	55.09%
NON-ENGLISH	80.70%	78.13%	77.64%
ABSTRACT	62.19%	63.72%	57.51%
IN VITRO	48.59%	39.94%	39.37%
OTHER	19.48%	24.20%	18.57%
REFS CHECKED	67.20%	66.19%	66.01%
NO CONC	42.64%	34.91%	36.89%
MODELING	51.68%	41.55%	44.24%
NO SOURCE	71.76%	52.61%	58.07%
METHODS	80.91%	39.63%	52.18%
NO EFFECT	20.25%	23.26%	17.03%
FOOD	32.83%	45.70%	37.56%
YEAST	66.24%	84.24%	73.12%
PUBL AS	79.01%	58.07%	66.11%
NO DURATION	80.48%	48.23%	58.35%
OGICAL TOXICANT	53.29%	28.55%	30.34%
NO TOXICANT	67.09%	60.06%	55.30%
Macro average	56.81%	54.66%	51.92%
Weighted average	73.70%	64.04%	62.31%

Table 2: Fuzzy Kappa scores for human- and model-screened Exclusion Reasons on 4427 new abstracts. Model performance is already on par with some human screeners. Work is ongoing to further improve performance for future datasets.

0			0		
Screener 1	Screener 2	Screener 3	Screener 4	Screener 5	MODEL
1	0.674	0.565	0.601	0.963	0.425
	1	0.559	0.644	0.769	0.470
		1	0.651	0.792	0.511
			1	0.638	0.552
				1	0 502

Figure 2: Using the extensive database of manually screened data also improved efficiency of binary inclusion/exclusion prediction. (A) Baseline model saves users 50% screening effort on average.





(B) Augmenting standard model with pretrained model via transfer learning provides additional benefits (mean mprovement of 9.5% WSS over the baseline prioritization model, but several datasets had significantly larger gains)

(C) 95% Recall estimate was accurate lower bound in majority of 75 test cases.



SWIFT-Active Screener for EcoTox

SWIFT Ac	TIVESCREENER EcoTox Project	€		ф -	占 brian.howarc
creen Refer	rence				Add New Review
Currently Scree	ening: Level 1 - Title & Abstract		1.9%		Inclusion Color Exclusion Color
	nte-Carlo-derived insights into dose-k ande, D. D., Nahum, A. E.; Physics in Medicine ar		relationships for 50 keV-2	25 MeV photon beams in w	
Abstract: The relationships between D, K and K-col are of fundamental importance in radiation dosimetry. These relationships are critically influenced by secondary electron transport, which makes Monte- Carlo (MC) simulation indispensable; we have used MC codes DOSRZnrc and FLURZnrc. Computations of the ratios D/K and D/K-col in three materials (water, aluminum and copper) for large field sizes with energies from 50 keV to 25 MeV (including 6-15 MV) are presented. Beyond the depth of maximum dose D/K is almost always less than or equal to unity and D/K-col greater than unity, and these ratios are virtually constant with increasing depth. The difference between K and K-col increases with energy and with the atomic					
difference be number of th megavoltage simple analyti given voxel to energy in this 0.5R(csda)((E) secondary ele MC-derived vo aluminum. Ar	-	y and with the atomic brium' small creasing field size. A ance 'upstream' from a crons depositing their pproximate to he mean initial agree well with 'exact' eV for water and esented and evaluated	Exclusion Reason	xcluded, why? S	
			Save and Next		Display Instructions
SWIFT ACTIVESCRE	EENER ECOTOX Project	BoscExc 🔍	🕒 🗅 🗠 📼 🏟		💄 brian.howard
eports					Add New Review
ort Type otox Exclusion Reaso	n Repor 💉				463 References
ID 🔦	Title	Au	thors	Screening Status	Model Generated Answer
933737	1H and 13C NMR spectral studies of conformatio	Singha, Netai C., Sathyanarayana, D. N.		Not Screened	Chem Methods
933738	Development of competitive enzyme-linked imm	Abad-Fuentes, A., Esteve-Turrillas, F. A., Agullo, C. et al.		Not Screened	Chem Methods
933739	Biological degradation of 4-chlorobenzoic acid by	Adebusoye, S. A.		Not Screened	Bacteria
933740	Degradation of 2,5- and 3,4-dichlorobenzoic acid	Adebusoye, Sunday A., Adeosun, Olumuyiwa A., Olofinlade, Bolanle B.		Not Screened	Bacteria
933741					
	Characterization of multiple chlorobenzoic acid-d	Adebusoye, S. A., Miletto, M.		Not Screened	Bacteria
933742	Characterization of multiple chlorobenzoic acid-d Influence of chlorobenzoic acids on the growth a	Adebusoye, S. A., Miletto, M. Adebusoye, S. A., Picardal, F. W., Ilori, M. C). et al.	Not Screened Not Screened	Bacteria
933743	Influence of chlorobenzoic acids on the growth a	Adebusoye, S. A., Picardal, F. W., Ilori, M. C). et al.	Not Screened	Bacteria
933743 933744	Characterization of multiple novel aerobic polych	Adebusoye, S. A., Picardal, F. W., Ilori, M. C Adebusoye, S. A., Picardal, F. W., Ilori, M. C). et al.). et al.	Not Screened Not Screened	Bacteria
5933742 5933743 5933744 5933745 5933746	Influence of chlorobenzoic acids on the growth a Characterization of multiple novel aerobic polych Growth on dichlorobiphenyls with chlorine substi	Adebusoye, S. A., Picardal, F. W., Ilori, M. C Adebusoye, S. A., Picardal, F. W., Ilori, M. C Adebusoye, S. A., Picardal, F. W., Ilori, M. C Adebusoye, S. A., Picardal, F. W., Ilori, M. C). et al.). et al.	Not Screened Not Screened Not Screened Not Screened	Bacteria Bacteria Bacteria

The screens above show several of the enhancements made to SWIFT Active Screener (Howard 2020, doi:10.1016/j.envint.2020.105623) to support literature curation for EcoTox. In A) we can see an abstract presented to the user for screening. Articles are prioritized for review using a deep learning neural network. For excluded articles, an exclusion reason is suggested by the computer and supporting words and phrases are highlighted. The system also includes several custom reports B) created in support of the EcoTox literature review process.

Conclusions

• EcoTox Active Screener uses Deep Learning to:

- Save an additional 9.5+% screening time (above baseline 50%)
- Accurately predict exclusion reasons (60-80% F1 score for common reasons)
- Explain its predictions using attention-highlighting

• The system is being piloted at EPA, and several refinements are planned.

For further information Please contact ruchir.shah@sciome.com <u>www.sciome.com</u>

