Revisiting Hierarchical Text Classification: Inference and Metrics

Roman Plaud [1,2] Matthieu Labeau [1] Antoine Saillenfest [2] Thomas Bonald [1]

[1]Institut Polytechnique de Paris [2]Onepoint

TL;DR

- Quantitative evaluation of HTC methods using hierarchical metrics and rigorous methodology.
- Introduction of Hierarchical WikiVitals (HWV), a high-quality HTC dataset from Wikipedia with a deep hierarchy.
- Extensive experiments on four HTC datasets including HWV, with a novel loss function and BERT model, achieving competitive results.

Hierarchical Text Classification

- Multilabel text classification with labels belonging to a predefined known hierarchy tree.
- Labels are not independent due to parent-child relationships.

Evaluation Methodology

Requirements for a metric in HTC to be a “good” metric :

- Does not depend on the hierarchical framework.
- Must take into account the severity of the errors [1].
- Must come with a inference methodology or be decoding-free.

Inference methodology = method to produce binary predictions from a probability distribution.

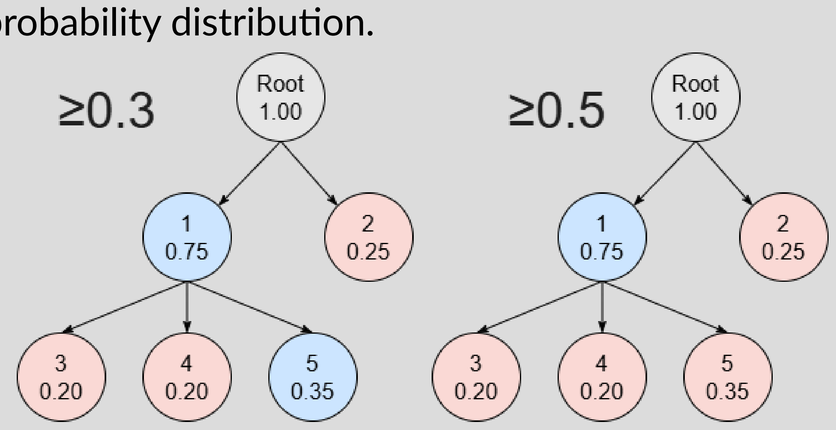

Figure 1. Two inference methodologies: thresholding to 0.3 (left) and 0.5 (right). Blue nodes correspond to predicted labels.

Identified metric : **AUC Hierarchical F1-score**

$$\mathrm{hP}(\mathrm{Y},\hat{\mathrm{Y}}) = \frac{\left|\hat{Y}^{\mathrm{aug}} \cap Y\right|}{\left|\hat{Y}^{\mathrm{aug}}\right|} \qquad \mathrm{hR}(\mathrm{Y},\hat{\mathrm{Y}}) = \frac{\left|\hat{Y}^{\mathrm{aug}} \cap Y\right|}{|Y|}$$

From Precision-Recall curve, area under the curve (AUC) is computed.
Decoding-free : takes into account the whole probability distribution.

Logit-adjusted Conditional Softmax

A novel loss function based on a conditional modelisation and whose objective is to adjust logits of a given class based on its frequency in the dataset (inspired by [2])

$$\hat{\mathrm{P}}(y|x,\pi(y)) = \frac{e^{s_x^{[y]} + \tau \log \nu(y|\pi(y))}}{\sum_{z \in \mathcal{C}(\pi(y))} e^{s_x^{[z]} + \tau \log \nu(z|\pi(z))}} \qquad \mathcal{L}(x,Y) = -\sum_{y \in Y} \log \hat{\mathrm{P}}(y|x,\pi(y))$$

- **Pros**: Deal with label imbalance and integrate prior hierarchy distribution.
- **Cons** : Only fitted for single-path leaf label.

Hierarchical WikiVitals: A novel dataset

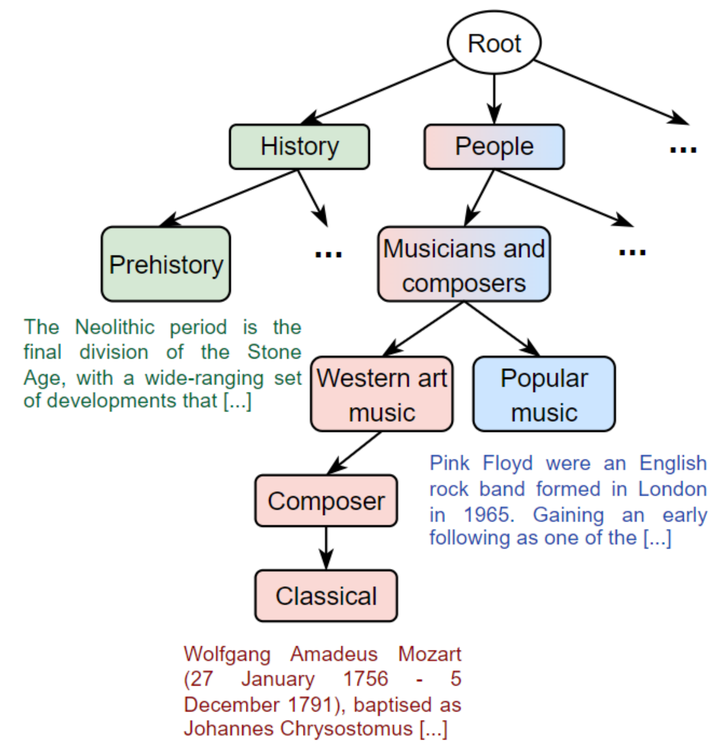

Figure 2. Extract of the taxonomy of our new dataset Hierarchical WikiVitals.

- Texts extracted from the abstracts of Wikipedia articles with handmade high labeling quality.
- Higher number of nodes (1186) and deeper hierarchy (depth of 6) compared to classical benchmark datasets.
- Imbalanced label distribution and ∼50% of labels have less than 10 examples in the dataset.

Results

	HWV	WOS	RCV1	BGC
BCE	$89.23_{\pm 0.13}$	$89.18_{\pm 0.10}$	$\mathbf{93.66}_{\pm 0.19}$	$\mathbf{90.26}_{\pm 0.29}$
CHAMP	$89.87_{\pm 0.19}$	$88.74_{\pm 0.08}$	$93.12_{\pm 0.33}$	$\mathbf{90.19}_{\pm 0.22}$
HBGL	–	$89.00_{\pm 0.10}$	$\mathbf{93.35}_{\pm 0.14}$	$88.08_{\pm 0.10}$
HGCLR	$88.35_{\pm 0.35}$	$\mathbf{89.23}_{\pm 0.22}$	$93.27_{\pm 0.14}$	$89.81_{\pm 0.17}$
HITIN	$90.72_{\pm 0.16}$	$88.92_{\pm 0.04}$	$93.04_{\pm 0.24}$	$\mathbf{90.08}_{\pm 0.16}$
Leaf Softmax	$88.55_{\pm 0.47}$	$88.62_{\pm 0.08}$	–	–
Conditional Sigmoid	$90.40_{\pm 0.17}$	$88.78_{\pm 0.17}$	$93.23_{\pm 0.36}$	$90.07_{\pm 0.40}$
Conditional Softmax	$\mathbf{90.94}_{\pm 0.09}$	$88.77_{\pm 0.07}$	–	–
Cond. Softmax + LA (ours)	$\mathbf{90.97}_{\pm 0.05}$	$88.90_{\pm 0.10}$	–	–

Table 1. hF1 AUC (and 95% confidence interval) on the test sets of the HWV, WOS, RCV1, and BGC datasets for the implemented models. Best results for each metric are highlighted in bold.

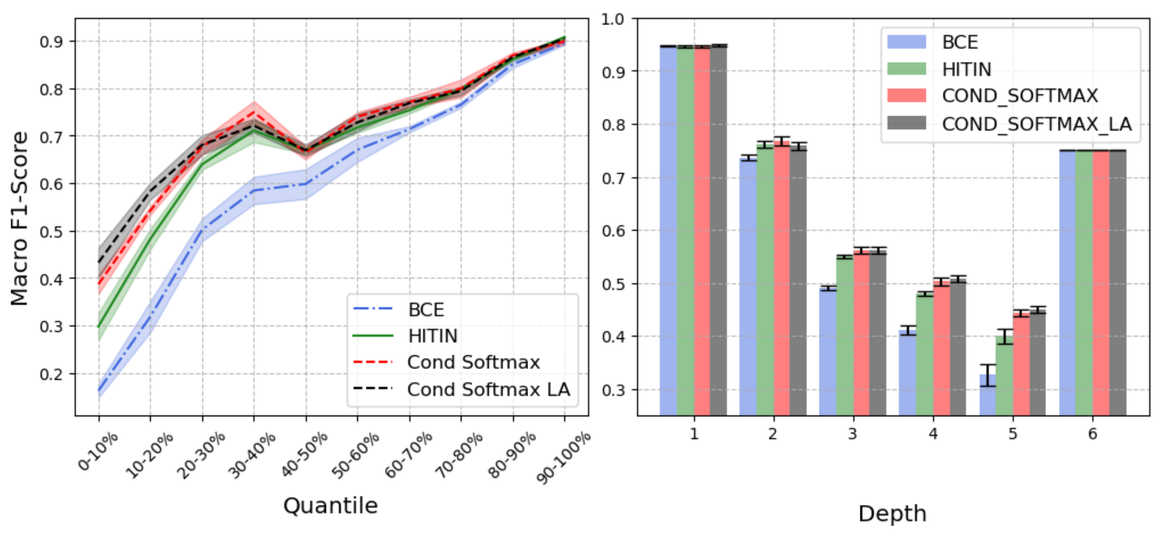

Figure 3. Averaged Macro F1-Scores on test set per quantile of the training set label distribution (left) and per depth (right) for different models and for the HWV dataset. The shaded regions (left) and error bars (right) represent a 95% confidence interval

Take Home Message

- Use adapted evaluation metrics with appropriate inference methodologies.
- Common datasets can be too simplistic; consider more challenging datasets like HWV.
- With proper evaluation, simple loss-based methods (e.g., logit-adjusted conditional softmax cross-entropy) can perform competitively with recent complex state-of-the-art models.

References

[1] Enrique Amigo and Agustín Delgado. Evaluating extreme hierarchical multi-label classification. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5809–5819, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[2] Aditya Krishna Menon, Andreas Veit, Ankit Singh Rawat, Himanshu Jain, Sadeep Jayasumana, and Sanjiv Kumar. Long-tail learning via logit adjustment. In International Conference on Learning Representations (ICLR) 2021, 2021.

contact : roman.plaud@telecom-paris.fr