About code availability: We have contacted the program chairs if we can provide source codes in the response, and they tell us that the constraint of 1-page PDF without external links applies to everyone. This means that we are unable to provide the source codes at this stage, due to the constraint of the paper reviewing procedure. However we will definitely submit the source code together with the final version of the paper. Please note that this point was also promised during the process of paper submission.

To reviewer 1

1. We will revise the paper carefully by incorporating the suggestions to make the paper clearer.

2. The superiority of AttentionXML comes from two key points: 1) multi-label attention, which captures the most important parts of texts for each label, allows to represent a given document differently for each label. In fact the ablation analysis in Section 3.6 shows that using BiLSTM instead of CNN improves the performance, which is however still much worse than Parabel. On the other hand, using multi-label attention instead of pooling makes the performance the best, outperforming Parabel, Bonsai and DiSMEC. Additionally, in Section 4 of Appendix, we show a typical example to demonstrate the advantage of attention mechanism in AttentionXML. 2) A shallow and wide PLT makes AttentionXML to handle extreme scale data efficiently.

3. The good performance of AttentionXML over tail labels can be attributed to the following two factors. 1) a shallow and wide PLT. In contrast to the deep balanced PLT with a large cluster size in Parabel, the PLT constructed in AttentionXML has a smaller tree height and cluster size. In this case, the chance of grouping unrelated (dissimilar) tail labels into one cluster (meta-label) is very small, which makes the model training of tail labels much easier and more accurate. Tables 2 and 3 in Appendix illustrate the impact of different height H and cluster size M on performance. 2) multi-label attention. Previous methods, such as Parabel and DiSMEC, used only one document representation for all labels, including many unrelated tail labels. It is difficult to satisfy so many unrelated (dissimilar) labels by the same text representation. As we discussed above, multi-label attention can handle this point effectively.

To reviewer 2

1. We will add a table on the notations used in our paper, revise the method section and add the pseudocodes on tree building, model training and prediction to make the paper clear and easy to follow.

2. Yes, in fact, our solution is like "a kind of additional negative sampling". By using such additional negative sampling, we can get a more precise approximation of log likelihood than only using nodes with positive parents. We will add detailed discussion on this point.

3. Models are trained and predicted sequentially from top to bottom. After training the current model for level i, we generate the candidates for level $i+1$. The value of parameter C we used in training is the same as the one used in inference. We will emphasize these details in the final version.

4. We will include the result of ExtremeText as another baseline. For clarity, the outputs of AttentionXML $\hat{y}$ correspond to the node variable z_n. We also use binary cross-entropy loss (Sorry for the typo of missing "binary").

To reviewer 3

1. We will add the three references suggested by the reviewer and ProXML as a competing method in experiments with respect to $PSP@k$ (we have contacted the authors of ProXML and are running ProXML with their generous help. Running ProXML is relatively time consuming, but we can put its result in the final version).

2. Following the reviewer's suggestion, we have examined the performance of AttentionXML under three settings (with shallow tree, deep tree and without PLT (No PLT: standalone attention mechanism)) on three relatively small datasets (Table 1). Note that "No PLT" is equivalent to a tree of only the root with L leaves and we used $K = M = 4$ for the other two settings. The experimental results showed that AttentionXML achieved the best performance under "No PLT" (without PLT) on all these three datasets. Also the performance decreased slightly with a deeper tree on all these datasets. This result implies that PLT is an approximation for achieving better scalability, losing the predictive accuracy slightly. Overall we think that this experiment highlights 1) the importance of multi-label attention mechanism to achieve high accuracy, and 2) that of PLT to achieve model scalability.

Table 1: Performance comparisons ($P@5$) of AttentionXML with different H. $H = 0$ means without a PLT.

AttentionXML	H	EUR-Lex	Wiki10-31K	AmazonCat-13K
No PLT	0	**61.10**	**68.78**	**66.90**
Shallow	2	60.88	67.27	66.28
Deep	4	60.54	65.89	65.46