AAAI 23 review

Overall

8 4 4 3

雪智：重点回4，掰3，不太管1（统一回123）

To 1 2 3

1: 用4的话回他，comlicated是有必要的，提出新坑让别人踩

```
prefer：之后我们会在各个小点上（Concept Tagging等）深挖，提出新的工作（画饼）
Or：或者说我们之后appendix给concept tagging的工作细节
Or：有标注和其他技术方案concept tagging，接上一行
```

举小的例子说明concept怎么来的，但是我们这篇只关注怎么用，他的来源有业务和打标（很杂，另一片paper才能讲清楚）希望后面有更多的人做我们这个方向（画饼）

Reviewer #1

{Strengths and Weaknesses}

The authors propose a two-stage framework TaLR, where the retrieval stage is based on vector similarity and concept information, and the reranking stage via a scoring engine with contrastive information. Experiments are conducted three datasets with more than 3 layers of taxonomy. Authors show that their method outperforms previous SOTA. The writing of this paper is clear.

My concern about the paper is that (1) the proposed method is a little heuristic and complicated (2) the authors should compare with general hierarchical text classification methods (both supervised and **zero-shot methods**).

{Questions for the Authors}

1. Would the authors like to compare their methods with some hierarchical text classification methods (general and product ones)?

E.g.,

[1] Hierarchy-Aware Global Model for Hierarchical Text Classification. ACL 2020.

[2] Weakly-Supervised Hierarchical Text Classification. AAAI 2019.

[3] Improving hierarchical product classification using domain-specific language modelling. WWW 2021.

[1] HiAGM是HiMatch的前序工作，我们已经比较了HiMatch（SOTA）再比较HiAGM的意义不是很大

[2] 专注于解决的两个challenge 1. 训练数据缺乏 2. 分类节点层级不好确定 （maybe不太适合我们的问题）

[3] 的贡献相当于验证了domain-specific continue pretraining的有效性

(OVERALL EVALUATION)

Borderline reject

Reviewer #1

{Strengths and Weaknesses}

The retrieval step is a combination of vector-based retrieval and rule-based concept-empowered retrieval. In my point of view, if the authors conduct knowledge-enhanced pretraining, vector-based retrieval alone might be enough. Besides, although the taxonomy is evolving all the time, we can always concatenate the name of the ancestor nodes with the name of the current node, to enrich the semantic meaning of the current node.

The authors didn't review related literatures on XMC (extreme multi label classification) http://manikvarma.org/downloads/XC/XMLRepository.html

There is limited novelty from the methodology perspective.

{Questions for the Authors}

What's the end-to-end model performance if we don't have concept-empowered retrieval?

(OVERALL EVALUATION)

Borderline reject

Reviewer #9

{Strengths and Weaknesses}

Strengths:

The paper describes a highly sophisticated system that likely powers an industrial e-commerce system. It is a good look at the engineering behind such systems. A good ablation study is provided as well and the paper feels rigorous. We also get two data sets of product-taxonomy labels which are presumably going to be useful for the community.

Weaknesses:

The biggest weakness I feel with this paper is that it is very hard to come away with any lessons after reading it.

The paper feels just like a list of experiments that might not hold up well outside the specific environment.

For instance, a significant amount of space is dedicated to explaining why concepts are a vital component and yet the concept tagger is only mentioned in a hand-wavy way: "The tagging step $X \rightarrow \{\lambda 1, \lambda 2, ...\lambda k\}$ is accomplished by an industrial Label Tagging System that exploits hybrid approaches including text sequence labeling, classification, literal matching and some expert-defined rules.".

这部分的解释见开头

So as a reader this is at best a brain-dump of ideas but without clarity on what lessons one can take from this work to future attempts at solving similar problems.

{Questions for the Authors}

I do not have too many questions. I would appreciate an example to clarify the method described in section 3.3. and the sampling strategy in 3.4.

I would say that these two sections are overly verbose.

For instance I see this statement: "This setting is tailored for the DMPC problem pursuing cross-domain alignment and uniformity, where inter-concept semantics are tied closer and intra-concept ones are further distinguished."

-> but nowhere in this paper are concepts being embedded - the retrieval stage uses the concept for some frequency computation only.

So would like a small example for these two stages.

(OVERALL EVALUATION)

Reject

照顾她（maybe改分）

Reviewer #11

{Strengths and Weaknesses}

Strengths:

1. The proposed approach outperforms strong baselines, and each component of the proposed approach is ablated, showing it's importance.

2. New dataset for product categorization in Chinese langauge will be released

Weaknesses:

1. The paper is not very well organized: - I am missing some examples that better define the problem, what is the input, what is the output, etc.

- The proposed method should be better described. I found Figure 2 hard to follow, as it is not clear where is the entry point of the "pipeline". Moreover, arrows looked a bit confusing -- they only point to the outside of the "title" box, also there is a disconnect between the title and "concept sets", etc. - The negative sampling procedure should be explained with more details.

- The "footnotes" under the tables are confusing, I would suggest the authors to incorporate them into the table caption itself. - Table 3 is confusing

2. The dataset should be better analyzed and presented -- this is an important contribution of the paper, so it should be clearly stated the language, now this ca only be inferred from the model type used for experiments. Further the paper should include analysis of the vocabulary, text lengths, overlaps, etc.

雪智：建议很好，回他我们之后会改

3. The results would be hard to reproduce, as important details about the models' hyperparameters and their training are missing.

4. The authors should report more details about the annotation process -- how many annotators, what's their agreement, what were the annotation guidelines, etc.

{Questions for the Authors}

1. Why two different metrics are used for the model -- accuracy (for the "Overall" results) and F1 for per-dataset? **Would it also make sense** to report ranking metrics such as AP, MRR, etc., in order to illustrate how far off your model is after the ranking?

雪智：基于什么考虑用了acc，解释一下

2. How are the hyperparameters optimized, is there a standard validation set that you used? There is no validation set size stated in Table 1.

(OVERALL EVALUATION)

Borderline reject

Reviewer #12

{Strengths and Weaknesses}

Pros:

- DMPC is common in e-com business. However, the research on it has been limited. This work is timely and will help the community to focus on this important task.

- Using a retrieval and ranking method to tackle DMPC is novel. In addition, the authors also smartly use concepts and contrastive learning to further improve their TaLR framework. - Experiments are comprehensive. Providing some case studies and measurement on running time are very helpful to industry practitioners.

Cons:

- Some details of the model and its training need more clarifications.

- Writing and graph illustration needs more revisions to make idea-convey clearer.

{Questions for the Authors}

1. Section 3.2 Negative sampling, why not using ground truth labels to form training data? Why k-fold m-class classifier setup is needed?

2. Figure 2, I cannot get the meaning of the two tree structures in the CL box.

3. Section 3.5 It is not clear to me what is BERT ranker's input. Do you concatenate all candidate labels to be an input to BERT or work on each label individually?

4. Regarding 3, which models are trained is not very clear. How many BERT models did you train? Did you use an end-to-end training or not?

5. In Table 2, some surprising patterns appear but have not been explained. - on the BH set, HiMatch-BERT is worse than BERT. This seems show that using hierarchy information does not help on the BH data. Any explanation? - On the FG set, why adding Concept always drags down performance for both BERT and HiMatch models?

雪智：不能怠慢这个reviewer，一条条回

(OVERALL EVALUATION)

Strong Accept