Achieving Diversity and Relevancy in Zero-Shot Recommender Systems for Human Evaluations

Tiancheng Yu[*]
MIT
yutc@mit.edu

Yifei Ma
AWS AI Labs
yifeim@amazon.com

Anoop Deoras
AWS AI Labs
adeoras@amazon.com

Abstract

Recommender systems (RecSys) often require user-behavioral data to learn good preference patterns. However, the user data is often collected by a working RecSys in the first place. This creates a gap where we hope to establish general recommendation patterns without relying on user data first, while the performance is then evaluated by real human oracles. On top of that, we aim to introduce diversity in the recommendation results, based on uncertainty principles to yield good trade-offs between recommendation coverage and relevancy.

Assuming that we have a corpus of item descriptions for all the items in our recommendation catalog, we propose two methods based on pretrained large language models (LLMs): Bert Corpus Tuning (Bert-CT) and Bert Variational Corpus Tuning (Bert-VarCT). Here, Bert-CT is responsible for adapting Bert to attend to domain-specific word tokens in the corpus of the item descriptions and Bert-VarCT is used to introduce diversity without significant changes in the network designs. We show that both methods achieved our designed goals, measured by data from real humans on a crowd-sourcing platform. Additionally, our approach is general and minimalistic. We release our codes for reproducibility and extensibility at `https://github.com/awslabs/crowd-coachable-recommendations`

1 Preliminaries

We consider the *item-to-item similarity-based recommendation* task. Let $\mathcal{I}$ be the set of available items. For each item $i \in \mathcal{I}$, let x_i be the associated textual description.[2] Given an item j which is last browsed by a user, we want to choose an item $i \in \mathcal{I}$ that is similar to item j. We will call the item j to be the source item in the following. An *recommendation agent* (or agent for short) $\mathcal{A}$ takes the source item j as input and outputs one candidate item $\mathcal{A}(j) \in \mathcal{I} \backslash \{j\}$ as the recommendation. Multiple recommendations in a slate are independently generated without replacements.

We use cosine similarity to retrieve relevant items with similar textual embeddings. Let $z = f_{\mathcal{A}}(x)$ with $\|z\| = \mathrm{Const.}$ be the encoder function from raw description texts to item embeddings. Notice that we append layer normalization to the encoder functions such that the output embeddings will always have equal norms. The retrieval function rounds up being the same as maximum inner-product search: $\mathcal{A}(j) = \mathrm{argtopk}_{i \in \mathcal{I} \backslash \{j\}} f_{\mathcal{A}}(x_i)^T f_{\mathcal{A}}(x_j)$. We consider two embedding functions which empirically worked well in our preliminary studies:

1. TF-IDF (term frequency-inverse document frequency, [1]). The embedding is a sparse vector whose dimension is equal to the vocabulary and each coordinate is the frequency of a term appearing in this description normalized by the number of descriptions where the term is present.

[*]Work completed during internship at AWS AI Labs.

[2]For product recommendations, textual descriptions contain much detailed information such as brand, sizes, and ingredients. We find image data to be mostly redundant in our experiments.

NeurIPS 2022 Workshop on Human in the Loop Learning. Dec 2, 2022. New Orleans, USA.

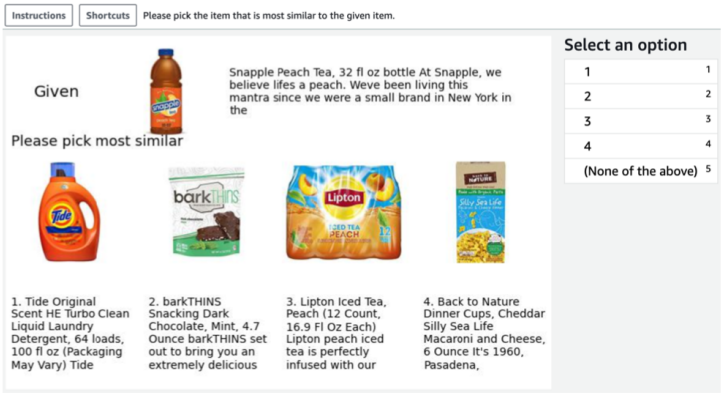

Figure 1: Demo recommendation application based on Amazon Review dataset. Here we show an evaluation setup where we shuffle the results from four different algorithms for comparisons by human annotators on AWS SageMaker GroundTruth platform.

2. Neural networks. In particular thanks to the sequential structure of the description x, we also consider transformer-based models such as Bert encoder [2]. The output z is usually a dense vector whose length is equal to the size of the summary layer of hidden states in a neural network.

Finally, we use Amazon Review dataset (specifically, the Prime Pantry subcategory) [3] as a running example throughout the paper. We ignore the user behaviors in this dataset because they contain confounding factors that lead to biased recommendations toward everyday products such as laundry detergents. Instead, we collect a set of ground truth similarity patterns by posing four item choices from four distinct algorithms (TF-IDF, pre-trained Bert, a model trained from the user-behavioral data provided by the original dataset, and a uniform-random sampler). See Figure 1 for an example of our data collection interface, which was run on AWS SageMaker GroundTruth platform. We evaluate the performance of an algorithm by asking it to rerank the four given options to check whether the top recommendation agrees with the human choice.

2 Zero-Shot Relevancy through Corpus Tuning (Bert-CT)

We first discuss our strategy to yield good relevancy in zero-shot recommendation settings. While the naive solution according to our preliminary is to utilize pre-trained large language models (LLMs) as encoder functions followed by cosine similarity search, our key observation is that we may obtain significantly better results if we fine-tune these LLMs on the target domain of all item descriptions.

Specifically, we use Bert masked training procedure where the (masked) input is pass through an encoder $q_\phi(z|x)$ to produce an item-level embedding z. We then use the decoder $p_\theta(x|z) = \prod_k p_\theta(w_k|z)$, where $x = [w_1, \ldots, w_k]$, to reconstruct the masked-out word tokens in the description of the same item. For simplicity, the target tokens are represented as a bag of words and predicted repeatedly from the same item embedding z, using a combination of linear projections and softmax activation. See Figure 2 for an illustration of the architecture. With 15% masking rate represented by the masking function $m(x)$, the training objective is:

$$\max_{\theta,\phi} \log p_\theta(m(x)|z), \text{ where } z = q_\phi(z|\tilde{m}(x)). \tag{1}$$

For empirical evaluation, Table 1 shows the performance of the proposed method and baseline algorithms. Here, Bert uses the inner-product of the raw hidden states of the Bert masked language model, while Bert-LayerNorm adds a layer-norm to keep the output to be vectors of constant norm. Bert-CT is our proposed method after domain adaptation and Bert-VarCT is a variational extension that we will introduce in Section 3. Compared with raw Bert models, the two proposed approaches have closed most of the gap toward TF-IDF.

Finally, we briefly discuss our intuitions behind Bert-CT. This is inspired by the strong empirical results of TF-IDF heuristics, where the IDF term contains summary statistics of the entire item

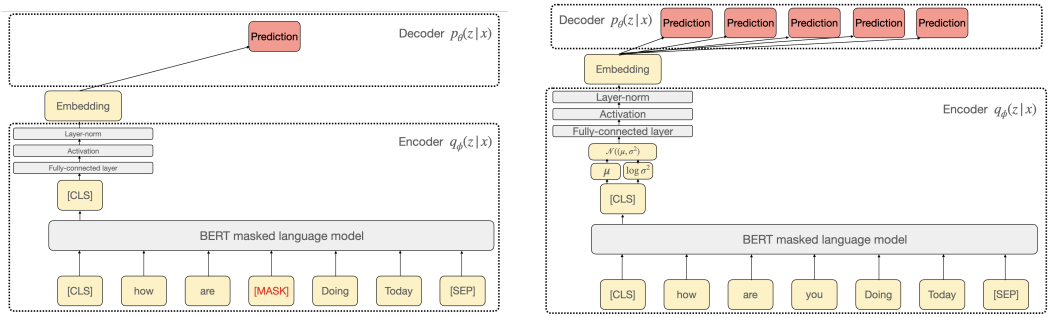

Figure 2: An illustration of the architecture of Bert-CT (left) and Bert-VarCT (right).

Table 1: Test of recommendation relevance by asking each agent to rerank a set of four recommendation items that have previously been shown to human annotators. Reporting the rate of agreements between the Top-1 recommendation from the agents and the original human annotations. Note that about 20% of the tasks were originally labeled for "none of the above", which lowers the average scores by the same factor for all agents. The error bounds could be estimated as $2/\sqrt{npq} \leq 0.050$.

Random	Bert	Bert-LayerNorm	Bert-CT	Bert-VarCT	TF-IDF	BM25
0.205	0.278	0.300	0.364	0.364	0.371	0.364

description corpus. We hypothesize that fine-tuning Bert models on the target domain yields a similar effect. In fact, we may take an analogy between the attention layers in the Bert models and the TF term. Through fine-tuning, we may discount the attention weights to common words found in the entire corpus. Notice that using LLMs learned from similar domains could also improve the zero-shot retrieval performance, but [4] suggests that transferring a similar-domain LLM does not significantly outperform simple heuristics like BM25. Our proposed Bert-CT remains a viable alternative for its extensibility to diversity goals, which we describe next.

3 Zero-Shot Diversity through Variational Corpus Tuning (Bert-VarCT)

Besides zero-shot recommendation relevance, another key property in zero-shot scenarios is *exploration* when we consider the opportunity for human feedback after our initial recommendations. We hope to make the candidate items in the choices not only relevant but also diverse so that the agent can learn something useful through the answers.

A simple heuristic is greedy exploration, where for the given source item j and the agent $\mathcal{A}$ we want to train, we choose item i with the highest scores $f_{\mathcal{A}}(x_i)^T f_{\mathcal{A}}(x_j)$.

A variant of greedy exploration, ϵ-greedy exploration, chooses a candidate randomly with probability $\epsilon < 1$ and greedily otherwise. This simple remedy does not work well because the randomness is not generated in the most informative direction. In other words, the trade-off between relevance and diversity is not efficient.

To address this issue, we use the Thompson sampling [5, 6, 7] framework, which is a principled Bayesian exploration rule and has strong theoretical guarantees [8].

In our problem setup, we implement the Thompson sampling framework with the following steps:

1. For every item i, estimate the posterior distribution of the embedding vector given the input text $q_\phi(z_i|x_i)$;
2. Draw embedding vectors $z_i \sim q_\phi(\cdot|x_i), \forall i$ and find $\text{argmax}_{i \in \mathcal{I} \setminus \{j\}} z_i^T z_j$ for one recommendation;
3. Repeat Step 2 to fill all recommendation positions without replacements.

In the above procedure, our main contribution is to provide distributional embedding $q_\phi(z|x)$ to model the uncertainty due to missing or uninformative item descriptions.[3] We call this model Bert-VarCT, or VarCT for short, inspired by Variational Auto-Encoder (VAE) framework. Note that there is a slight abuse of notations where the old function $q_\phi(z|x)$ in the Bert-CT method could be seen as the mean value of the new definition of $q_\phi(z|x)$, which is a distribution in the Bert-VarCT approach.

The intuition behind Bert-VarCT is based on a Bayesian principle where we start with a prior model with large variance $p_0(z)$ and we use the observed data, i.e., the item description data, to adjust the posterior estimate via Bayes rule, $p_\theta(z|x) \propto p_0(z)p_\theta(x|z)$. This typically results in variance reduction when more description data is observed. Conversely, the posterior distribution $p_\theta(z|x)$ preserves a larger uncertainty when the item descriptions are missing or uninformative.

We use a standard ELBO (evidence lower-bound) derivation to approximate the posterior model via a feed-foward *variational* encoder network $q_\phi(z|x)$. The derivation below highlights the relationship between the original Bayesian integral likelihood (left side) and our ELBO likelihood (right side). We use ELBO for its computational benefits.

$$\begin{aligned}\max_{\phi,\theta} \log \mathbb{E}_{p_0(z)}[p_\theta(x|z)] &= \log \mathbb{E}_{q_\phi(z|x)}\left[\frac{p_0(z)p_\theta(x|z)}{q_\phi(z|x)}\right] \\ &\geq \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{\mathrm{KL}}(q_\phi(z|x)||p_0(z)) \end{aligned} \tag{2}$$

We further simplify $q_\phi(z|x)$ and $p_0(z)$ as Gaussian distributions using a common reparametrization trick. See Figure 2 for the architectural details of Bert-VarCT and notice the correspondence between $q_\phi(z|x)$ and $p_\theta(x|z)$ in the equations and their positions in the diagram.

In our implementation, we add a tuning parameter β to quantify the trade-off diversity and relevance, inspired by β-VAE [9].[4] Our modified ELBO is $\mathbb{E}_{q_\phi(z|x)} \log p_\theta(x|z) - \beta D_{\mathrm{KL}}(q_\theta(z|x)||p_0(z))$. The larger the β, the stronger the regularization becomes, and the more diversity we can expect from the trained models.

Figure 3 shows different sets of recommendations due to different random seeds from the same recommendation algorithm. We notice that the Greedy-CT method keeps generating the same set of recommendations due to the lack of randomness. The ϵ-greedy agent generates different questions each time, but the candidates are either the same as the greedy agent or totally irrelevant since they are randomly drawn from the pool. In this regard, our proposed VAE agent is superior at diversified candidate generation, where the candidates are also relevant to the given source item most of the time.

To make a more quantitative comparison, we measure the diversity by the number of unique items in ten consecutive recommendations. For each model, we also evaluate its precision and plot them in Figure 4. Here the CT+noise curve is generated by adding IID Gaussian noise in the hidden state, to introduce some randomness in the CT agent. Compared with the ϵ-greedy agent baseline, both CT+noise and VarCT agent achieves significantly higher precision for a given level of diversity or a much higher diversity for any given level of precision. As result, the experiment shows that VarCT (and its simplified version, CT+noise) have the power to achieve an efficient trade-off between diversity and relevance.

Besides the positive results, we also notice a limitation, where VarCT did not produce significant improvements over CT+noise as expected. We hypothesize that this may be due to a shallow decoder design (we used iid distribution instead of more complex text-generation models) and the fact that our task might be too simple. Nonetheless, these limitations do not prevent us from using VarCT to generate diversity in the recommendation choices and we leave model tuning to future work.

4 Conclusion and future work

In this paper, we study two key challenges in the problem of zero-shot recommendations for human evaluations, namely relevancy and diversity. We propose Bert-CT to close the gap between pre-trained Bert recommenders and heuristic TF-IDF agents; and we propose VAE-based model to yield

[3] We omit the subscript i in the rest of the section.

[4] Another motivation for β-VAE is through the principle of maximum entropy under the constraint of reconstruction quality [10]. See [11, 12] for further discussions as well as a proposal for better selections of the prior distributions. Similar principles have also been used for robotic explorations [13].

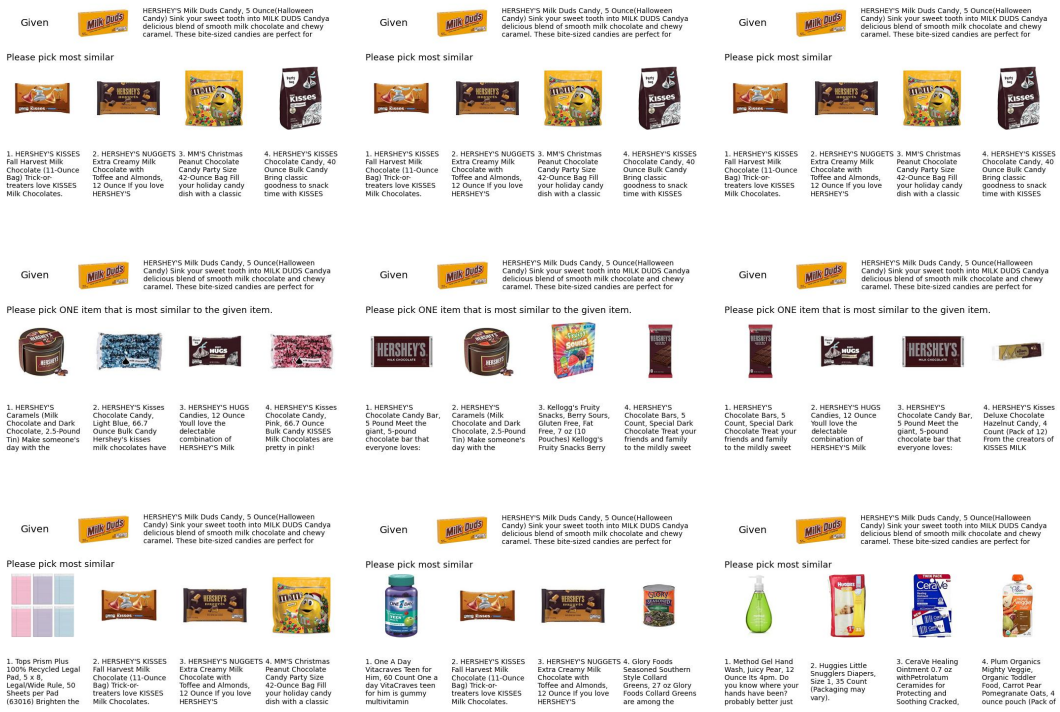

Figure 3: Each row is three multiple choice questions generated by one agent in three recommendations. Row 1: Greedy. Row 2: VAE with $\beta = 0.002$. Row 3: ϵ-greedy with $\epsilon = 0.5$.

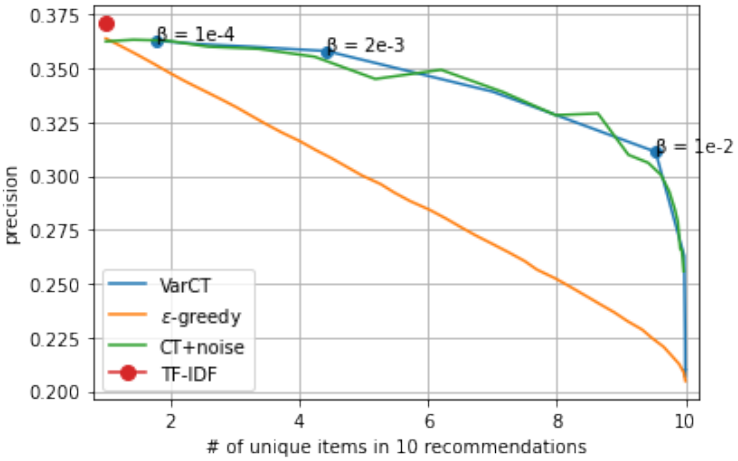

Figure 4: Diversity vs relevancy for different agents with different parameters. The x-axis is the number of unique items appearing in ten independent recommendations and the y-axis is the precision evaluated on the prime pantry reranking task. Each agent will correspond to a point in the plane and for both axis, the higher the better. Since there will be an intrinsic trade-off between diversity and precision, we can only expect the trade-off the be more smooth for better agents.

significantly better diversity with least compromises in relevancy. For future work, we hope to close the loop where we utilize the collected human feedback for further improvements. We expect that the uncertainty principles we adopted for diversity generation could also imply efficient data collection strategies due to their close connections to Bayesian information gain.

Acknowledgements

We thank Ge Liu and Andrey Kan for early discussions of ideas. We also thank Stefano Soatto for editorial feedback as well as references for the additional discussions on the properties of β-VAE.

References

[1] Juan Ramos et al. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*. New Jersey, USA, 2003.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[3] Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 188–197, 2019.

[4] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[5] William R Thompson. On the theory of apportionment. *American Journal of Mathematics*, 57(2):450–456, 1935.

[6] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

[7] Daniel J Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, Zheng Wen, et al. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.

[8] Shipra Agrawal and Navin Goyal. Further optimal regret bounds for thompson sampling. In *Artificial intelligence and statistics*, pages 99–107. PMLR, 2013.

[9] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2016.

[10] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

[11] Alessandro Achille and Stefano Soatto. Information dropout: Learning optimal representations through noisy computation. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2897–2905, 2018.

[12] Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *The Journal of Machine Learning Research*, 19(1):1947–1980, 2018.

[13] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. Maximum entropy inverse reinforcement learning. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, 2008.