Interpretable Word Representation Learning Framework for Modeling Semantic Relevance in E-commerce

Haoyue Zhang [⋆1], Zhiyuan Zeng[∗2], Guoliang Zhang[1], Hui Zhao[2], Tianshu Wu[2], Pengjie Wang[2], Jian Xu[2], Bo Zheng[2] ✉, and Baolin Liu[1] ✉

[1] University of Science and Technology Beijing, China
[2] Alibaba Group, Beijing, China
{zhanghaoyue,zhangguoliang}@xs.ustb.edu.cn, {liubaolin}@ustb.edu.cn
{zengzhiyuan.zzy,shuqian.zh,shuke.wts}@alibaba-inc.com
{pengjie.wpj,xiyu.xj,bozheng}@alibaba-inc.com

Abstract. In e-commerce search, the semantic relevance model assesses the relevance between user search queries and candidate product titles to determine which products can be presented to users. To balance efficiency and accuracy, in industrial scenarios, most current methods adopt relevance models based on late-interaction. However, these methods do not effectively model the fine-grained relevance between query and title, especially for long-tailed words. In this paper, we propose the Interpretable Word Representation Learning Framework, a novel relevance model in e-commerce, which overcomes this challenge and improves the accuracy of relevance calculation. Specifically, to enhance the model's ability to determine the semantic relevance of word pairs, we propose the sparse alignment based on optimal transport (OT), which focuses on one-way semantic alignment from query to title when calculating relevance. To enhance the model's ability to represent long-tail words, we combine character-level and word-level semantic representations to encode word information. In addition, we design an effective contrastive learning method to train our model to alleviate the problem of insufficient negative samples. Extensive experiments have verified the excellent performance of our proposed method. We have successfully deployed it to our online search system.

Keywords: E-commerce · Semantic Relevance · Representation Learning · Optimal Transport

1 Introduction

Semantic relevance modeling in e-commerce can be viewed as a problem of calculating the relevance between a user's query and product titles. Previous methods include Bi-encoders [3, 9], which have limited accuracy due to the lack

[⋆] Haoyue Zhang and Zhiyuan Zeng contributed equally to this work.
✉ Baolin Liu and Bo Zheng are the corresponding authors.

of semantic internal interaction. The cross-encoder [9], whose interaction layer considers global information, has a high computational cost. The current solution is the late-interaction models [11,17], which better captures semantic relationships through the late interaction mechanism and achieves a good balance between accuracy and efficiency. However, these methods are still insufficient in modeling fine-grained semantic relevance, resulting in two challenges that limit the ability of e-commerce relevance models. First, they fail to effectively model the word alignment relationships between queries and titles, resulting in weak ability to determine the relevance of word pairs. Second, since these methods ignore character-level information, they perform poorly in representing long-tail words.

For the first challenge, we propose the sparse alignment method based on OT [4] that emphasizes the one-way word alignment from query to title. Studies have shown that finding cross-sentence alignment helps analyze sentence similarity [6]. In e-commerce relevance tasks, we focus on whether each intent in the query is satisfied in the title. Based on experience and statistics, each word in the query usually conveys a specific intent independently. Therefore, for each word in the query, we consider aligning it with the word in the title that are most semantically similar to it and emphasize the contribution of these word pairs to the overall relevance. This method better explains and measures the similarity and enhances the model's determination of fine-grained semantic relevance. We use OT theory to explain this one-way alignment through two matrices: the cost matrix, which represents the word alignment score between query and title; and the alignment matrix, which represents the word alignment relationship.

For the second challenge, traditional relevance models focus on word-level modeling while ignoring character-level information, making it difficult to learn accurate semantic representations of low-frequency words or out-of-vocabulary words through limited samples [5]. To address this, we propose the multi-level encoder that combines character-level and word-level representations. The encoder consists of two components: the word-level encoder that processes character sequences and the sentence-level encoder that further processes word representations. This encoder architecture can represent long-tail words more accurately.

Furthermore, we adopt an effective contrastive learning method to alleviate data sparsity and high annotation cost issues in industrial scenarios [7, 14]. We use users' click data to construct positive query-title sample pairs. Then we generate negative pairs through both random in-batch sampling and hard negative sampling. In summary, the main contributions of our work are as follows:

- We propose the sparse alignment method based on OT, which improves the model's ability to calculate the semantic relevance of word pairs.
- We propose a novel encoder modeling approach, the multi-level Encoder, which enhances the model's representation of long-tail words by encoding both character-level and word-level information.
- Extensive experiments and analyses validate our method's superior performance in e-commerce relevance task. We have successfully deployed this approach on our online search system, improving the Relevant rate by 2.07%.

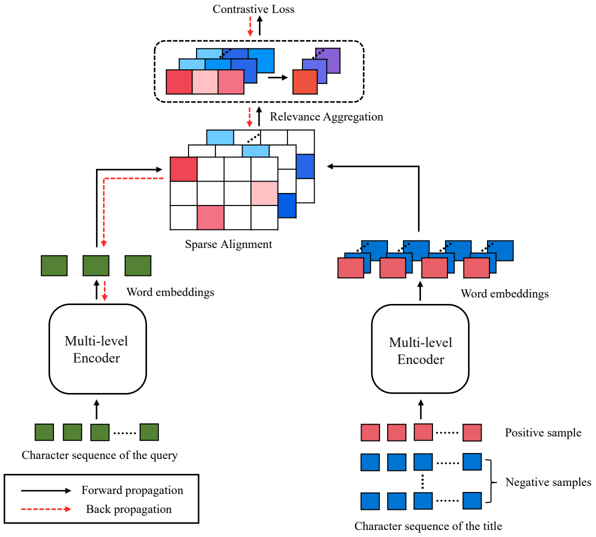


Fig. 1: The architecture of proposed contrastive learning relevance model.

2 Related Work

Text Matching Problem. Current text matching methods can be mainly divided into three categories. Bi-encoders, such as DSSM [3] and Sentence-BERT [9], use two independent encoders to process the query and the document separately, followed by computing similarity between their vectors. The second is cross-encoder [9], which combine the query and the document into a single input, capturing interactions comprehensively at a high computational cost. The late-interaction models [8], such as the classic ColBERT series [11,17], encode query and the document at the token level and uses a late-interaction module to compute relevance scores to balance accuracy and efficiency.

Optimal Transport in NLP. Optimal transport (OT) is used to measure the distance between two distributions and has been widely applied to NLP tasks [1,2,6]. For example, applying OT theory to word alignment can effectively solve the problem of word frequency imbalance and improve the accuracy and interpretability of alignment. [6] uses OT to measure the distance between sentences. [2] develops new monolingual word alignment methods to solve the problems of null alignment and one-to-many alignments.

3 Approach

Figure 1 shows our approach's overall framework. Given a query q and a title t, we encode them into $\boldsymbol{E}_q = \{\boldsymbol{E}_{q_1}, \boldsymbol{E}_{q_2}, ..., \boldsymbol{E}_{q_m}\}$, and $\boldsymbol{E}_t = \{\boldsymbol{E}_{t_1}, \boldsymbol{E}_{t_2}, ..., \boldsymbol{E}_{t_n}\}$, using the multi-level encoder proposed in Section 3.2, where $\boldsymbol{E}_{q_i}, \boldsymbol{E}_{t_j} \in \mathbb{R}^d$. We then calculate the relevance between q and t using the method in Section 3.1.

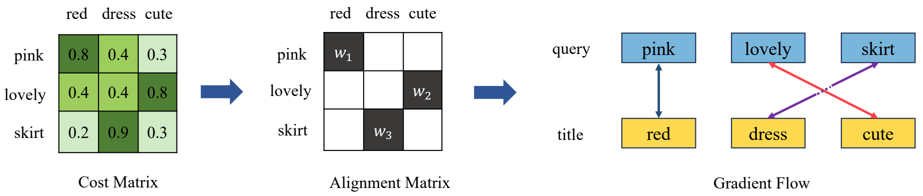


Fig. 2: An example of the sparse alignment method based on OT, where the word alignment relationship is obtained through the cost matrix and alignment matrix.

3.1 Sparse Alignment Based on Optimal Transport

Sentence Relevance as an Optimal Transport Problem. Optimal Transport Problem (OT) aims to identify the most efficient method of mass transfer between two probability distributions to minimize their transfer cost, involving transfer strategies and costs [4]. We use OT to interpret the sentence similarity problem, which is to seek the least costly method to transfer elements between these two sentences by calculating the cost matrix and alignment matrix.

For the pair of q and t, the cost matrix $\boldsymbol{C} \in \mathbb{R}^{m \times n}$ represents the distance between any word pairs of the two sentences. The alignment matrix $\boldsymbol{T} \in \mathbb{R}^{m \times n}$ signifies the proportion of transmission between pairs of words. The alignment matrix $\boldsymbol{T}$ shows how each word in q should be aligned to a word in t. Then the similarity between two sentences can be expressed as follows:

$$L_C(q,t) := \min \sum_{i,j} \boldsymbol{C}_{i,j} \boldsymbol{T}_{i,j}, \tag{1}$$

Where $\boldsymbol{C}_{i,j}$ and $\boldsymbol{T}_{i,j}$ represent the transmission cost and transmission ratio between q_i and t_j, respectively. This method measures sentence relevance in more detail and increases the interpretability of sentence relevance in e-commerce.

Semantic Sparse Alignment Relations of Words. Now we describe how to get the $\boldsymbol{C}$ and the $\boldsymbol{T}$ between a query and a title. As shown in Figure 2, we first get $\boldsymbol{C}_{i,j}$ by calculating the cosine similarity of word embeddings $\boldsymbol{E}_{q_i}, \boldsymbol{E}_{t_j}$ in q, t, and then obtain the alignment matrix $\boldsymbol{T}$ based on $\boldsymbol{C}$.

In order to determine whether each intent in the query is satisfied, based on experience and actual statistics on e-commerce data, we only need to focus on the one-way transmission from query to title. Therefore, for each word in the query, we find the word in the title that is semantically closest to it, that is, the one with the maximum cosine similarity value. Then we assign a transportation ratio $w_i \in (0,1)$ to the position corresponding to this maximum value in $\boldsymbol{T}$. Each row of $\boldsymbol{T}$ has only one non-zero position, indicating the sparse alignment relationship. Then, the $\boldsymbol{C}$ and $\boldsymbol{T}$ can be represented as:

$$\boldsymbol{C}_{i,j} = \cos(\boldsymbol{E}_{q_i}, \boldsymbol{E}_{t_j}) \tag{2}$$

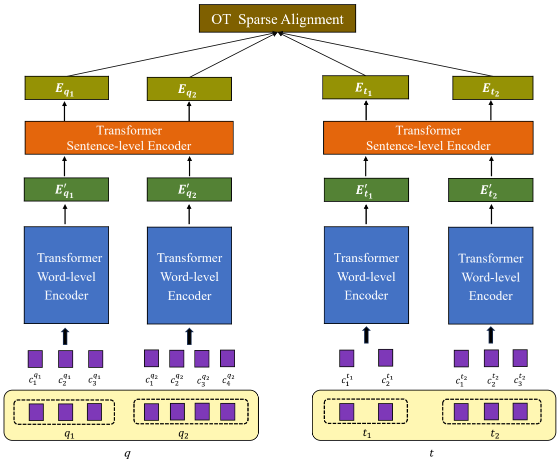

OT Sparse Alignment
E_{q_1}
E_{q_2}
E_{t_1}
E_{t_2}
Transformer Sentence-level Encoder
Transformer Sentence-level Encoder
E'_{q_1}
E'_{q_2}
E'_{t_1}
E'_{t_2}
Transformer Word-level Encoder
Transformer Word-level Encoder
Transformer Word-level Encoder
Transformer Word-level Encoder
$c_1^{q_1}$ $c_2^{q_1}$ $c_3^{q_1}$
$c_1^{q_2}$ $c_2^{q_2}$ $c_3^{q_2}$ $c_4^{q_2}$
$c_1^{t_1}$ $c_2^{t_1}$
$c_1^{t_2}$ $c_2^{t_2}$ $c_3^{t_2}$
q_1
q_2
t_1
t_2
q
t

Fig. 3: The overall structure of the multi-level encoder, including word-level encoders and sentence-level encoders.

$$\boldsymbol{T}_{i,j} = \begin{cases} w_i & if\ \boldsymbol{C}_{i,j} = \max(\boldsymbol{C}_{i,1}, \boldsymbol{C}_{i,2}, \dots, \boldsymbol{C}_{i,n}) \\ 0 & otherwise \end{cases} \tag{3}$$

Since $\boldsymbol{T}$ is a sparse matrix, the similarity between q and t is determined by aggregating the maximum values of each row in $\boldsymbol{C}$, using weights w_i from $\boldsymbol{T}$. The calculation method for w_i will be detailed in Section 3.3.

$$sim(q,t) = \sum_i \boldsymbol{C}_{i,j} w_i, \text{ where } j = \arg\max_{j'} \boldsymbol{C}_{i,j'} \tag{4}$$

3.2 The Multi-level Encoder

For the encoder, we propose the multi-level encoder that integrates character-level and word-level information, as shown in Figure 3. We use character-level information as model input, and the encoder includes the word-level encoder f_W and the sentence-level encoder f_S, both of which are based on Transformer [13].

Word-level Encoder. For a pair of inputs q and t, we tokenize each word into character-level tokens $c_1^{q_i} c_2^{q_i} ... c_k^{q_i}, c_1^{t_j} c_2^{t_j} ... c_l^{t_j}$. We use word-level encoder f_W to process the input character sequences. Specifically, we use several parameter-sharing f_Ws to encode each word in q and t, and f_W consists of multiple identical Transformer encoder blocks. After obtaining the embedding of each token, the embedding of each word $\boldsymbol{E}'_{q_i}, \boldsymbol{E}'_{t_j}$ is obtained through the average pooling operation $Pooling_{MEAN}$, where $\boldsymbol{E}'_{q_i}, \boldsymbol{E}'_{t_j} \in R^d$.

$$\begin{aligned} \boldsymbol{E}'_{q_i} &= Pooling_{MEAN}(f_W(c_1^{q_i}, c_2^{q_i}, ..., c_k^{q_i})), \text{and} \\ \boldsymbol{E}'_{t_j} &= Pooling_{MEAN}(f_W(c_1^{t_j}, c_2^{t_j}, ..., c_l^{t_j})) \end{aligned} \tag{5}$$

Sentence-level Encoder. We further encode q and t using two parameter-sharing sentence-level encoders f_Ss, each of which consists of several connected Transformer encoder blocks[1]. The preliminary word vectors $\boldsymbol{E}'_{q_i}, \boldsymbol{E}'_{t_j}$ obtained previously are used as the input of the two f_Ss. By considering the context information, the final word vectors $\boldsymbol{E}_{q_i}, \boldsymbol{E}_{t_j}$ are generated for each word.

$$\begin{aligned} \boldsymbol{E}_{q_i} &= f_S(\boldsymbol{E}'_{q_1}, \boldsymbol{E}'_{q_2}, ..., \boldsymbol{E}'_{q_m}), \text{and} \\ \boldsymbol{E}_{t_j} &= f_S(\boldsymbol{E}'_{t_1}, \boldsymbol{E}'_{t_2}, ..., \boldsymbol{E}'_{t_n}) \end{aligned} \tag{6}$$

Then the similarity between q and t is calculated using the sparse alignment method based on OT proposed in Section 3.1.

3.3 Contrastive Learning Method for E-commerce Relevance Learning

Data Construction. We construct positive query-title sample pairs based on users' high Click-Through Rate (CTR) behaviors. For negative samples, we randomly select titles within the training batch and manually create category and attribute negatives. Category negatives share the same category but differ in attributes (e.g., brand, color), while attribute negatives belong to different categories but share some attributes with the center title. The ratio of category to attribute negatives is approximately 4:1.

Training Objective. We first introduce the specific calculation of the similarity between q and t. For each word in q, we define the maximum value of each row in $\boldsymbol{C}$ introduced in 3.1 as the relevance score r_i of the word:

$$r_i = \max_j \boldsymbol{C}_{i,j} \tag{7}$$

To satisfy the user's search intent, we should pay attention to each word in the query, and focus on the intent in the query that is not satisfied in the title. Then the smaller r_i is, the larger its corresponding aggregation weight w_i is, which is the value in $\boldsymbol{T} \in \mathbb{R}^{m \times n}$ introduced in 3.1. Specifically, we use the *softmin* to calculate w_i, which yields a value close to the minimum through a smooth, differentiable method. This is typically defined as the inverse of the *softmax* function, achieved by applying softmax after negating the input.

$$w_i = \frac{e^{-r_i/\tau}}{\sum_{i=1}^{m} e^{-r_i/\tau}} \tag{8}$$

Where τ controls the temperature. The similarity between q and t can be expressed as:

$$sim(q,t) = \sum_i w_i \cdot r_i \tag{9}$$

[1] In our work, f_W and f_S consist of 12 and 2 connected Transformer encoder blocks, respectively.

The contrastive learning training objective for $(q_i, t_i^+, t_i^-, t_i'^-)$ in a training batch of size N is as follows, where, t_i^-, $t_i'^-$ represent category negative samples and attribute negative samples respectively.

$$\ell_i = -\log \frac{e^{sim(q_i,t_i^+)/\tau}}{\sum_{j=1}^{N}(e^{sim(q_i,t_j^+)/\tau} + e^{sim(q_i,t_j^-)/\tau} + e^{sim(q_i,t_j'^-)/\tau})} \tag{10}$$

4 Experiment

4.1 Dataset

In our experiments, the dataset consists of one training set and two test sets.

Ecom-relevance Training Dataset comes from our online e-commerce platform. We use the method in Section 3.3 to construct 40 million positive sample pairs and 480 million negative sample pairs for training our model and all baseline models.

Annotated Dataset is a manually annotated dataset for e-commerce, consisting of 700,000 query-title pairs labeled as *relevant* or *irrelevant* by experienced annotators. Due to its limited number, it is used only as a test set.

EcomRetrieval [15] is a public dataset, derived from the C-MTEB(Chinese Massive Text Embedding Benchmark) dataset, used to assess Chinese text embedding models for retrieval tasks. It contains 1,000 queries and around 100,000 labeled product titles from real e-commerce scenarios.

4.2 Experimental Setup

Implementation details. For sample construction, we select query-title pairs with CTR values greater than 0.3 as positive sample pairs. We use Tensorflow to implement the model, using the Adam optimizer, a learning rate of 1e-5, a training batch size of 512, and 5 epochs of training. We use 8 A100 GPUs to train the models, and the average inference time is tested on a single V100 GPU. These hyperparameters are selected based on experimental experience.

Baselines. We compare our model with several popular types: BM25 [10] and Bi-encoders [3, 9, 15], such as BGE [15], a state-of-the-art text embedding model for retrieval and relevance tasks. Late-interaction models [6, 11, 15, 17], where ColBERT [17] uses the "MaxSim" interaction method to calculate relevance. We also evaluate the performance of the cross-encoder [9] and Qwen2-7B [16].

Evaluation Metrics. We validate the binary classification capability of the model on the Annotated Dataset and the ranking capability on EcomRetrieval. Therefore, we report the following metrics to fully evaluate the performance of the model: **AUC(Area Under Curve)**, **F1(F1-score)**, and **R@k(Recall)**.

4.3 Main Results

Table 1 shows the performance comparison of various methods. Late-interaction models balance accuracy and efficiency compared to BM25, classic Bi-encoders,

Table 1: Performance of our model compared with baselines.[2]

Methods	Inference time	Annotated Dataset		EcomRetrieval		
		AUC	F1	R@10	R@50	R@1000
Lexical systems:						
BM25 [10]	78ms	68.85	65.73	70.1	82.1	89.5
Bi-encoders:						
DSSM [3]	122ms	78.80	77.42	76.7	87.1	93.7
Sentence-BERT [9]	168ms	79.59	78.30	77.8	87.8	94.1
BGE [15]	170ms	80.77	78.56	79.4	89.1	95.1
Late-interaction:						
ColBERT [17]	283ms	80.92	78.85	80.6	89.8	95.5
ColBERTv2 [11]	265ms	81.51	79.87	82.3	91.0	96.4
CLRCMD-BERT$_{\text{base}}$ [6]	308ms	81.77	80.04	82.9	91.4	96.4
BGE-ColBERT [15]	288ms	81.80	80.35	82.8	91.6	96.5
Ours	299ms*	83.50*	81.72*	84.5*	92.8*	97.1*
Cross-encoder [9]	1577ms	83.91	82.24	85.6	93.6	97.5
Qwen2-7B [16]	11785ms	85.48	83.59	86.7	94.3	98.0

* The values of AUC, F1, and R@k in the table are expressed as percentages, with the percentage sign omitted.

and cross-encoder. Our method outperforms ColBERT series models in classification and ranking tasks, demonstrating precise semantic representation of queries and titles and effective relevance computation through its interaction mechanism. Compared with CLRCMD-BERT$_{\text{base}}$, employs a bidirectional OT method, our one-way word alignment approach from query to title shows better performance in e-commerce relevance tasks, effectively reducing incorrect semantic alignment from title to query.

In online e-commerce systems, we have strict limits on model inference time. We compared Qwen2-7B and a cross-encoder in experiments. Qwen2-7B generates 0/1 results to represent whether it is relevant or not. We use the probability of generating 1 at the corresponding position in the generated sequence to represent the relevance score. Although Qwen2-7B has higher accuracy, we cannot deploy LLM (large language model) online due to its long inference time. Our model achieves the best accuracy among the Bi-encoders and late-interaction models while ensuring efficiency, and is very close to the accuracy of the cross-encoder.

4.4 Ablation Study

In this section, we conduct ablation studies on each module. **w/o Sparse Alignment** directly calculates cosine similarity between the query and title for

[2] All pre-trained models in the baselines are fine-tuned using our training set, maintaining consistent experimental settings and comparable parameter sizes.

Table 2: Ablation study on Annotated Dataset and EcomRetrieval test datasets.

Methods	Inference time	Annotated Dataset		EcomRetrieval		
		AUC	F1	R@10	R@50	R@1000
Ours	299ms	**83.50**	**81.72**	**84.5**	**92.8**	**97.1**
w/o Multi-level Encoder	263ms	83.15	81.40	84.1	92.6	97.0
w/o Sparse Alignment	206ms	80.77	78.76	80.4	89.5	95.3
w/o Both	155ms	80.16	78.52	79.5	88.9	94.9
self-sup.	–	76.78	74.05	75.2	86.0	93.0
weakly sup.	–	80.98	78.81	81.0	90.0	95.5

* The values of AUC, F1, and R@k in the table are expressed as percentages.

the relevance score and its encoder is consistent with 3.2. As shown in Table 2, our results outperform **w/o Sparse Alignment** in all metrics, demonstrating the effectiveness of the sparse alignment module. Precise one-way word alignment and late interaction mechanisms enhance semantic similarity calculation between query and title. **w/o Multi-level Encoder** employs a 12-layer Transformer encoder followed by the sparse alignment module. Results show the multi-level encoder better represents semantic information and improves similarity computation. The results show that the multi-level encoder can better represent semantic information and calculate similarity.

Table 2 compares the inference time of different methods. We can find that although the sparse alignment module incurs a certain time consumption, the significant performance improvement it brings to e-commerce relevance tasks is well worth it for our online system.

We also conducted an ablation study on the data construction strategy. Using the fully self-supervised contrastive learning method **self-sup.** as the baseline, we found it struggles to adapt to complex e-commerce data distributions due to insufficient data diversity. The **weakly sup.** method with random negative sampling is too simple. Our results show that the two proposed negative sample construction methods significantly improve model performance.

4.5 Model Analysis

We perform ablation studies to analyze model components using data from our online e-commerce platform. For each word, we retrieve a synonym list ranked by predicted relevance scores, with manually annotated relevance labels.

The effect of Sparse Alignment. We further analyze the effectiveness of the sparse alignment component through experiments, as shown in Table 3. We selected 100 words and calculated the average MAP@k value (Mean Average Precision at top k) [12] based on their synonym lists. The results show that **Ours** is significantly better than **w/o Sparse Alignment** in terms of MAP indicators, verifying that considering word alignment in sentence relevance calculation helps the model determine the semantic relevance of word pairs.

Table 3: Ablation analysis of the effectiveness of different modules.

Methods	MAP@10(%)	MAP@20(%)	MAP@50(%)	$Dist_{high}$	$Dist_{tail}$
Ours	**87.82**	**83.24**	**78.54**	**0.874**	**0.838**
w/o Multi-level Encoder	82.80	79.11	77.02	0.840	0.704
w/o Sparse Alignment	75.45	70.37	67.56	0.757	0.710
w/o Both	67.79	65.31	63.93	0.732	0.616

The effect of Multi-level Encoder. Following we analyze the advantages of the multi-level encoder in representing long-tail words. The better the model represents long-tail words, the closer the distance between the center word and the set of synonymous long-tail words. We define the $Dist(\Gamma, v)$ between a set of words and a center word v as follows, by computing the average cosine similarity between each word in the set and the center word:

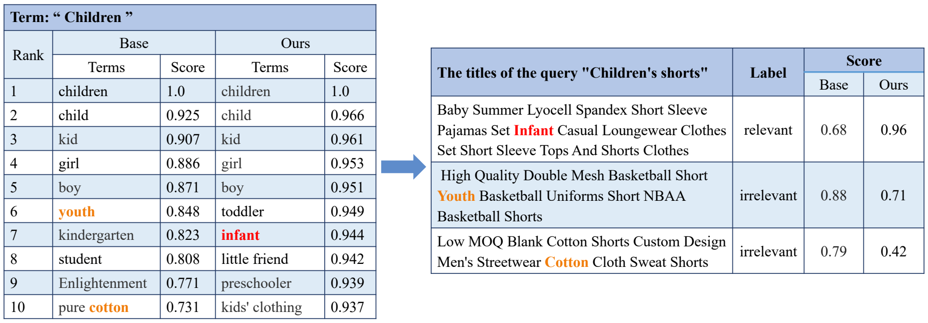
$$Dist(\Gamma, v) = \frac{1}{|\Gamma|} \sum_{t_i \in \Gamma} \cos(\boldsymbol{E}_{t_i}, \boldsymbol{E}_v) \tag{11}$$

Where $|\Gamma|$ denotes the number of words in the set Γ. We measure the average distances of 100 center words to synonymous high-frequency and long-tail word sets, $Dist_{high}$ and $Dist_{tail}$. Table 3 shows that our model more accurately calculates the relevance between synonym sets and the center word. Comparing **Ours** with **w/o Multi-level Encoder**, removing the multi-level encoder has little effect on high-frequency words but significantly lowers long-tail word accuracy from 0.838 to 0.704. The results between **w/o Sparse Alignment** and **w/o Both** also highlight the multi-level encoder's advantage for long-tail words. Comparing **w/o Both** and **w/o Multi-level Encoder**, sparse alignment module reduces the distance between long-tail words and the center word. However, adding the multi-level encoder in **Ours** brings the distance between the center word and long-tail words closer to that of high-frequency words, indicating that the multi-level encoder primarily enhances long-tail word representation.

4.6 Case Study

As shown in Figure 4, we analyze the better effect of our method in the e-commerce relevance task through specific examples. The baseline **Base** is ColBERT, the previous online relevance model, trained with the same data and settings as **Ours**. For a word in the query, we retrieve synonyms. Compared with **Base**, our method obtains more accurate word pairs relevance calculation results.

We further analyze the relevance of this query and its titles. For the title labeled *relevant*, **Ours** achieves a higher relevance score by accurately calculating the relationship between "*Children*" in the query and "*Infant*" in the title. For the *irrelevant* title, compared with the **Base** model, **Ours** accurately determines that "*Children*" is not relevant to "*Youth*" and "*Cotton*". This proves that our model can better learn semantics and determine the relevance of key word pairs.



Term: " Children "				
Rank	Base		Ours	
	Terms	Score	Terms	Score
1	children	1.0	children	1.0
2	child	0.925	child	0.966
3	kid	0.907	kid	0.961
4	girl	0.886	girl	0.953
5	boy	0.871	boy	0.951
6	**youth**	0.848	toddler	0.949
7	kindergarten	0.823	**infant**	0.944
8	student	0.808	little friend	0.942
9	Enlightenment	0.771	preschooler	0.939
10	pure **cotton**	0.731	kids' clothing	0.937

The titles of the query "Children's shorts"	Label	Score	
		Base	Ours
Baby Summer Lyocell Spandex Short Sleeve Pajamas Set **Infant** Casual Loungewear Clothes Set Short Sleeve Tops And Shorts Clothes	relevant	0.68	0.96
High Quality Double Mesh Basketball Short **Youth** Basketball Uniforms Short NBAA Basketball Shorts	irrelevant	0.88	0.71
Low MOQ Blank Cotton Shorts Custom Design Men's Streetwear **Cotton** Cloth Sweat Shorts	irrelevant	0.79	0.42

Fig. 4: A case analysis comparing our model and the Base model.

4.7 Online Evaluation

The proposed model has been integrated into the search advertising platform and has been operational for over a year. It exhibits an online Query Per Second (QPS) of 40,000. To handle a large number of user queries, we pre-compute and store all token vectors. During online serving, the system only needs to process a lightweight late-interaction computation.

To evaluate the proposed model, an online A/B test was conducted over 30 days, with each model receiving 5% of search traffic. The benchmark for comparison was the existing online relevance model, ColBERT. Both models assessed query-title relevance scores and filtered low-relevance pairs.

The outcomes revealed that the proposed model achieved a 0.41% uplift in Click-Through Rate (CTR), a 0.10% uplift in Revenue per Mille (RPM), a 0.32% increase in Conversion Rate (CVR), and a 0.58% boost in Return On Investment (ROI). Additionally, it improved the Relevant rate by 2.07%, which is used to evaluate the relevance of a user's search to the search results.

In summary, the online evaluation confirms that the proposed model outperforms its predecessors, significantly enhancing user experience for merchants and consumers while safeguarding the platform's revenue.

5 Conclusion

In this work, we construct a novel semantic relevance framework for e-commerce relevance tasks by introducing the sparse alignment based on OT, the multi-level encoder, and the contrastive learning method. It effectively models fine-grained semantic relevance, thereby improving the model's performance in relevance determination. Extensive experiments and analysis confirm the effectiveness of the proposed method, and we successfully applied it to the online search system.

6 Acknowledgements

This research was supported by the National Key Research and Development Program of China (No.2024YFC3308300), the National Natural Science Foundation of China (No.U2133218), the Alibaba Group's Research Institute Internship Program and the Fundamental Research Funds for the Central Universities of China (No.FRF-MP-19-007 and No. FRF-TP-20-065A1Z).

References

1. Alqahtani, S., Lalwani, G., Zhang, Y., Romeo, S., Mansour, S.: Using optimal transport as alignment objective for fine-tuning multilingual contextualized embeddings. arXiv preprint arXiv:2110.02887 (2021)
2. Arase, Y., Bao, H., Yokoi, S.: Unbalanced optimal transport for unbalanced word alignment. ArXiv **abs/2306.04116** (2023), https://api.semanticscholar.org/CorpusID:259095560
3. Huang, P.S., He, X., Gao, J., Deng, L., Acero, A., Heck, L.: Learning deep structured semantic models for web search using clickthrough data. Proceedings of the 22nd ACM international conference on Information & Knowledge Management (2013), https://api.semanticscholar.org/CorpusID:8384258
4. Kantorovich, L.V.: On the translocation of masses. Journal of Mathematical Sciences **133**(4), 1381–1382 (2006)
5. Kim, Y., Jernite, Y., Sontag, D., Rush, A.M.: Character-aware neural language models. Computer Science (2016)
6. Lee, S., Lee, D., Jang, S., Yu, H.: Toward interpretable semantic textual similarity via optimal transport-based contrastive sentence learning. arXiv preprint arXiv:2202.13196 (2022)
7. Liu, S., Liu, S., Xu, W.: Gated attentive convolutional network dialogue state tracker. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6174–6178. IEEE (2020)
8. Pang, L., Lan, Y., Guo, J., Xu, J., Wan, S., Cheng, X.: Text matching as image recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 30 (2016)
9. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks (2019)
10. Robertson, S.E., Zaragoza, H.: The probabilistic relevance framework: Bm25 and beyond. Found. Trends Inf. Retr. **3**, 333–389 (2009), https://api.semanticscholar.org/CorpusID:207178704
11. Santhanam, K., Khattab, O., Saad-Falcon, J., Potts, C., Zaharia, M.: Colbertv2: Effective and efficient retrieval via lightweight late interaction (2022), https://arxiv.org/abs/2112.01488
12. Shen, J., Qiu, W., Shang, J., Vanni, M., Ren, X., Han, J.: Synsetexpan: An iterative framework for joint entity set expansion and synonym discovery (2020), https://arxiv.org/abs/2009.13827
13. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. arXiv (2017)
14. Xiao, R., Ji, J., Cui, B., Tang, H., Ju, X.: Weakly supervised co-training of query rewriting andsemantic matching for e-commerce. ACM (2019)
15. Xiao, S., Liu, Z., Zhang, P., Muennighof, N.: C-pack: Packaged resources to advance general chinese embedding. arXiv preprint arXiv:2309.07597 (2023)
16. Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., Huang, F., et al.: Qwen2 technical report. arXiv preprint arXiv:2407.10671 (2024)
17. Zaharia, M., Khattab, O.: Colbert: Efficient and effective passage search via contextualized late interaction over bert (2020)