Advancing E-Commerce Search with KR-BERT and Entity Confidence

Alan Chuang
CS 274 - Web Intelligence
Prof. Teng Moh

Problem Definition

Context

- E-commerce platforms often struggle with accurately linking products to relevant concepts due to:
  - Ambiguous product descriptions by sellers
  - Product data is often noisy
  - Traditional LM’s lack structured, domain-specific knowledge.

Objective

- Enhance language representations to improve product-concept matching through Knowledge Relevance BERT (KR-BERT).
  - KG triplet embedding
  - Dynamic Relevance Scoring and Attention Mechanism

Stratified Sampling of Amazon Reviews Dataset

Amazon Reviews Dataset

- Categories: 33
- Reviews: 571.54M
- Users: 54.51M
- Items: 48.19M
- Gigantic dataset!!!


```
{'rating': 5.0,
 'title': 'Such a lovely scent but not overpowering.',
 'text': "This spray is really nice. It smells really good,
 'images': [],
 'asin': 'B00YQ6X8EO',
 'parent_asin': 'B00YQ6X8EO',
 'user_id': 'AGKHLEW2SOWHNMFQIJGBECAF7INQ',
 'timestamp': 1588687728923,
 'helpful_vote': 0,
 'verified_purchase': True}
```


Organize reviews into strata by:

- Product Categories
- Rating Levels (1-5 stars)
- Proportional sample size for each stratum.
- Conduct random sampling within each strata to get 10000 samples
- 10,000 x 5 (rating, title, text, item_id, user_id)

$$n_i = (N_i / N) \times n$$

n_i is the sample size for stratum i

N_i is the population size of stratum i

N is the total population size

n is the total sample size desired

Dataset Representation - Knowledge Graphs

What is a Knowledge Graph?

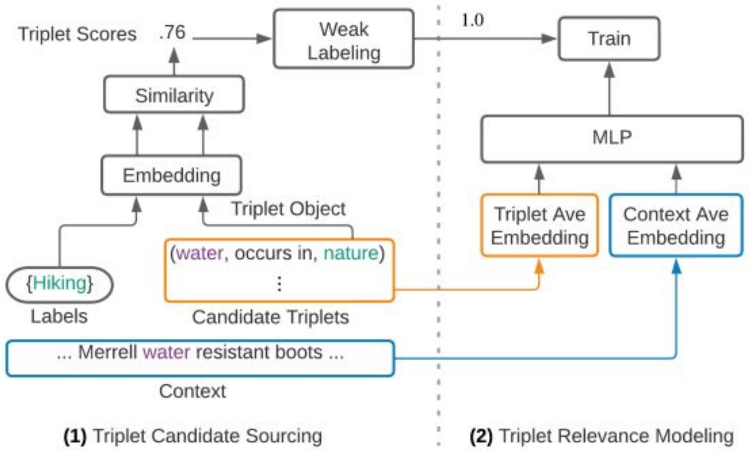
- Structured Network of entities (nodes) and relationships (edges) connecting them.

Application in KR-BERT Model:

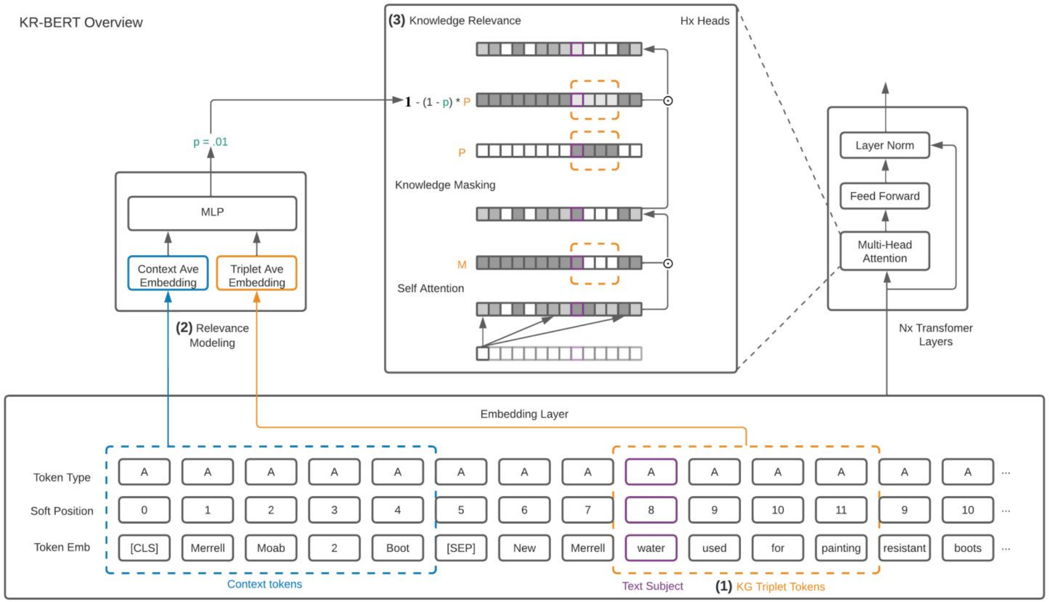
- Integrating knowledge graph data allows BERT to not only rely on the words' contextual usage but also on their relation to real-world entities and concepts.

Knowledge graph triplets:

- EX: [*Product - categorized as - Category*]



KR-BERT Architecture



Attention in BERT:

- The relative importance of different words (or tokens) in a sequence.

(1) Triplet Embeddings:

- Convert KG triplets into vector representations, included with word embeddings for relevance modeling.

(2) Relevance Modeling:

- Determines how relevant each KG triplet is to the text being processed.

(3) Processing and Output:

- Combined embeddings pass through self-attention, normalization, and a feed-forward network.
- Outputs contextualized embeddings from original text + KG relevance.

Baseline Implementations

My Implementation

Method	F1	P	R
KR-BERT (baseline)	.6767	.6971	.6944
KR-BERT (Confidence)	.8137	7208	.7864

Paper Results

Method	Relevance Model	F1	P	R
KR-BERT	(frozen)	.703	.673	.814
KR-BERT	(Lcls only)	.700	.696	.793
KR-BERT (proposed)	yes (Lcls + Lrel)	.717	.697	.826

Baseline Insights

Keyword Search:

- Least impacted since not heavily reliant on the volume/complexity of data.

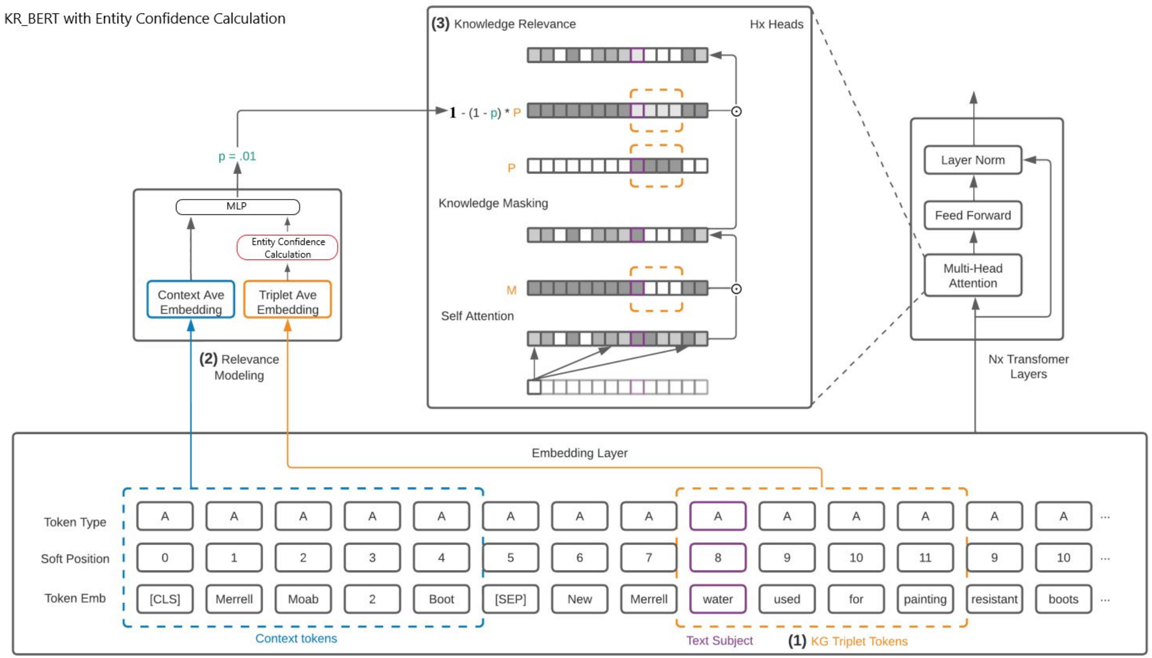
KG Lookup:

- Performed worse comparatively than Keyword Search, since new KG has reduced coverage.

KR-BERT:

- Surprisingly, performed almost similar to Keyword Search.
- Could be overfitting due to undersampling/oversimplifying the dataset.
- Noise in KG could also affect performance.

Proposed Architecture - Entity Confidence



Entity Confidence:

- Adjust attention weights based on the reliability of knowledge graph entities.

Components of Confidence Scoring:

- Frequency-Based Scoring:
  Entities that appear more frequently are more reliable.
- Context Consistency:
  How consistently entities appear in similar contexts.
- Source Reliability:
  How reliable is the entity source (e.g. verified purchase).

Confidence Score [0, 1] =

α × Frequency Score +
β × Context Score +
γ × Source Score

Proposed Solution Insights

Entity Confidence and Targeted Attention

- **Frequency-Based Scoring:**
  - KR-BERT will focus more attention on high confidence entities.
  - Prioritizes reliable information over less relevant data.
  - This in turn reduces noise in the overall dataset.
- **Stability:**
  - E-commerce data can vary greatly in quality and structure due to human input.
  - Entity confidence helps maintain consistency by choosing reliable entities.
  - By filtering out unreliable data points, the model is less likely to propagate errors from noisy data.

Analysis

Model	Avg Precision	Avg Recall	Avg F1
KR-BERT (original)	0.697	0.826	0.717
KR-BERT (baseline)	0.6971	0.6767	0.6944
KR-BERT (confidence)	0.7208	0.7864	0.8137

Key Insights:

- Enhanced data quality and contextual understanding with confidence scoring

- Stability through filtering unreliable data points

- Potential for further improvement with larger sample size and advanced preprocessing

Future Work

Increasing Sample Size:

- Expanding the sample size for training and evaluation can help our model to generalize better.

Extended Dataset Evaluation:

- Expanding the evaluation to other e-commerce categories and datasets to generalize the findings.

Refinement of Confidence Scoring:

- Further refining the confidence scoring mechanism by incorporating additional factors such as user trustworthiness and review helpfulness scores.

Real-time Adaptation:

- Implementing real-time adaptation of the model to continuously improve based on new data and user interactions.

Comparison with Other Models:

- New advanced models, such as GPT-4 and T5, could be used to benchmark performance and improve embedding generation.

Conclusion

- This study proposed enhancements to KR-BERT for improving product-concept matching on e-commerce platforms
- Integrating KG triplet embeddings and a dynamic relevance scoring mechanism significantly improved precision, recall, and F1 scores
- The inclusion of confidence scoring helped better handle noisy data and emphasize reliable information
- Insights highlight the potential of leveraging structured knowledge to enhance language models' performance in domain-specific applications
- Future work will focus on expanding evaluation, refining the model, and exploring real-world applications to maximize benefits of KR-BERT in the e-commerce domain

References

[1] K. Samel et al., “Knowledge Relevance BERT: Integrating Noisy Knowledge into Language Representations,” AAAI Conference on Artificial Intelligence, 2023.

[2] “McAuley-Lab/Amazon-Reviews-2023 · Datasets at Hugging Face,” Hugging Face, Mar. 31, 2024.

[3] S. Khalid, “BERT Explained: A Complete Guide with Theory and Tutorial,” Medium, Apr. 10, 2020.

[4] arrrrrmin, “arrrrrmin/albert-guide,” GitHub, Dec. 15, 2023.

[5] G. Singh, “Fine-Tune ERNIE 2.0 for Text Classification,” Medium, Aug. 2019.

[6] “How to Code BERT Using PyTorch - Tutorial With Examples,” neptune.ai, May 20, 2021.

[7] CheeKean, “Mastering BERT Model: A Complete Guide to Build it from Scratch,” Data And Beyond, Sep. 05, 2023.

[8] A. Dadoun, “Knowledge Graph Embeddings 101,” Medium, May 23, 2023.

[9] X. Ge, Y.-C. Wang, B. Wang, and C.-C. J. Kuo, “Knowledge Graph Embedding: An Overview,” arXiv.org, Sep. 21, 2023.

[10] N. Kolitsas, O.-E. Ganea, and T. Hofmann, “End-to-end neural entity linking,” arXiv preprint arXiv:1808.07699, 2018.

[11] W. Liu, P. Zhou, Z. Zhao, Z. Wang, Q. Ju, H. Deng, and P. Wang, “K-bert: Enabling language representation with knowledge graph,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 03, 2020, pp. 2901–2908.

[12] Q. Gu, Y. Zhang, J. Cao, G. Xu, and A. Cuzzocrea, “A confidence-based entity resolution approach with incomplete information,” 2014 International Conference on Data Science and Advanced Analytics (DSAA), 2014, pp. 97-103. doi:10.1109/DSAA.2014.7058058.