FolkScope: Intention Knowledge Graph Construction for Discovering E-commerce Commonsense

Changlong Yu¹, Weiqi Wang¹, Xin Liu¹, Jiaxin Bai¹, Yangqiu Song¹
Zheng Li², Yifan Gao², Tianyu Cao², Bing Yin²
¹The Hong Kong University of Science and Technology, Hong Kong SAR, China
²Amazon.com Inc, Palo Alto, USA
{cyuaq, wwangbw, xliucr, jbai, yqsong}@cse.ust.hk, {amzzhe, yifangao, caoty, alexbyin}@amazon.com

ABSTRACT
As stated by Oren Etzioni, “commonsense is the dark matter of artificial intelligence.” In e-commerce, understanding users’ needs or intentions requires substantial commonsense knowledge, e.g., “A user bought an iPhone and a compatible case because the user wanted the phone to be protected.” In this paper, we present FolkScope, an intention knowledge graph construction framework, to reveal the structure of humans’ minds about purchasing items on e-commerce platforms such as Amazon. As commonsense knowledge is usually ineffable and not expressed explicitly, it is challenging to perform any kind of information extraction. Thus, we propose a new approach that leverages the generation power of large language models and human-in-the-loop annotation to semi-automatically construct the knowledge graph. We annotate a large amount of assertions for both plausibility and typicality of an intention that can explain a purchasing or co-purchasing behavior, where the intention can be an open reason or a predicate falling into one of 18 categories aligning with ConceptNet, e.g., IsA, MadeOf, UsedFor, etc. Then we populate the annotated information to all automatically generated ones, and further structure the assertions using pattern mining and conceptualization to form more condensed and abstractive knowledge. We evaluate our knowledge graph using both intrinsic quality measures and a downstream application, i.e., recommendation. The comprehensive study shows that our knowledge graph can well model e-commerce commonsense knowledge and can have many potential applications.

1 INTRODUCTION
In e-commerce or online shopping platforms, understanding users’ searching or purchasing intentions can benefit and enable a lot of recommendation tasks [7, 13, 49]. Existing intention-based studies on recommendation systems are either of limited intention categories [7, 49] or using models to implicitly model the intention memberships [13]. In general large online e-commerce platforms, such as Amazon, acquiring or capturing users’ purchasing intentions in a scalable way is very challenging. Intentions are mental states where agents or humans commit themselves to actions. Understanding others’ behaviors and mental states requires rationalizing intentional actions [18], where we need commonsense, or, in other words, good judgements [26]. For example, “in a birthday party, we usually need a birthday cake.” Commonsense knowledge can be factoid [11] which is not invariably true. Meanwhile, commonsense knowledge is usually ineffable and not expressed explicitly. Thus, it is very challenging to acquire such kind of knowledge in existing text data. Thus, as an AI pioneer, Oren Etzioni, stated, “commonsense is the dark matter of artificial intelligence.” We still lack of technologies to explicitly discover the structure of our mind to rationalize our behaviors. Particularly, for users’ purchasing behaviors, there is still no scalable way to know such commonsense like “a user bought an iPhone and a compatible case because the user wanted the phone to be protected.”

Existing related knowledge graphs (KGs) can be categorized into two folds. First, there are some general situational commonsense KGs dealing with everyday social situations [36, 38, 52]. Such KGs focus on action consequences and cause/effect relations among events and states. However, they are not directly related to products on e-commerce platforms, so even the neutralized models, e.g., COMET [4], trained on such KGs cannot generalize well on users’ behaviors data. Another line of work leverages existing KGs, such as ConceptNet [26, 40] and Freebase [2], to integrate them into the e-commerce catalog data [30, 48, 53]. However, such an integration is still based on factual knowledge such as IsA and DirectorOf relations, and does not truly model the commonsense knowledge for purchase intentions. Recently, AliCoCo2 [29] attempted to tackle the commonsense extraction problem from question answering (QA) pairs, product descriptions, and reviews. They leveraged the SQuAD-style QA models [35] to answer questions such as “what kind of category has the functionality of terms?” where terms are some informal commonsense concepts. In this way, they are able to build a large scale KG about the factual knowledge more than product related concepts, e.g. space, crowd, time, function, and event, as indicated in ConceptNet [26].

However, existing KGs on e-commerce data can only indicate the plausibility of factual relations, and cannot fully reveal the intention of purchases. In fact, an intention acts as a mediator between action and what people believe and desire [22]. This can be reflected by the typicality of commonsense. For example, a user bought an iPhone 13 because “iPhone 13 has the function of taking photos” and “iPhone 13 can be used for social networking,” where the reasons can be plausible functions, whereas a more typical reason would be “the user’s previous phone is also an iPhone so it’s easy to transfer data,” “the users’ old phone is too slow,” or “the user is simply a fan of Apple products.” Thus, no matter what kind of factual knowledge a KG contains, if it is not directly linked to the rationalization, it cannot be regarded as typical commonsense. Furthermore, recommendation explanation has been proposed as a task to explain why a user rates high or low of an item [24, 31], where they use online reviews as the natural annotation of explanation. However, online reviews are noisy and diverse. Most of the time, reviews do not directly reflect the intention of the purchases, but rather the consequences of purchases or reasons of...
We divide the annotation process into two steps. In the first step, we can probe intentions in different perspectives mapped to the what typical reasons are for an additional item being bought to-

Then the more abstract intentions are formed to condense the representation of intentions. The intentions can be represented by arbitrary and loosely constrained, we also align our prompts with patterns or potential artifacts from human writers [25], which thus improves the diversity of knowledge.

Given generated candidates and annotations, to construct the KG, we first perform pattern mining to remove irregular generations. Then for both annotation steps, we train classifiers to populate the scores to all generated data. Finally, for each of the generated intentions, we perform conceptualization to map the key entities or concepts in the intention to more high-level concepts so that we can build a denser and more abstractive KG for future generalization. An illustration of our KG is shown in Figure 1. To verify the overall quality, we randomly sample the populated assertions to estimate the overall quality of the KG. We also use a downstream task, collaborative filtering based recommendation, to prove our KG’s quality and usefulness. The contributions of our work can be summarized as follows.

- We propose a new framework, FolkScope, to construct large-scale intention KG for discovering e-commerce commonsense knowledge.
We leverage large-scale language models to generate candidates and perform two-step large-scale annotation on Amazon data with two domains, clothing and electronics, and the process can be well generalized to other domains.

- We define the schema of the intention KG aligning with famous commonsense KG, ConceptNet, and populate a large KG based on our generation and annotation with 184,146 items, 217,108 intentions, 857,972 abstract intentions, and 12,755,525 edges (assertions).
- We perform a comprehensive study to verify the validity and usefulness of our KG. Both code and data will be public released for downstream tasks.

2 RELATED WORK

Knowledge Graph Construction. Here we briefly review general commonsense KGs and product related KGs. An early approach of commonsense KG construction is proposed in ConceptNet [26] where both text mining and crowdsourcing are leveraged. Later, ConceptNet is extended by integrating a lot of factual knowledge from other resources [40]. In 2012, a web-scale KG, Probase, which focuses IsA relations, is constructed based on pattern mining [46]. The advantage of this probabilistic KG is that it can model both plausibility and typicality of conceptualizations [39]. Recently, commonsense has attracted more attention in the field of AI and NLP. Particularly, Event2Mind [36] and ATOMIC [38], which are based on human annotations, are developed to model the situational knowledge about action consequences and cause/effect relations of events and states. Then their extensions and neuralized models are developed [4, 19]. Meanwhile, in KnowablyWood [41], WebChild [42], and ASER [51, 52], it is proven that information extraction can be used to extract event-related knowledge from dependency parsing and discourse analysis based on large-scale corpora. The extracted knowledge can then be transferred to other human annotated knowledge resources, e.g., ConceptNet [50] and ATOMIC [9].

In the e-commerce domain, Amazon Product Graph [48] is developed to align Amazon catalog data with external KGs such as Freebase and to automatically extract thousands of attributes in millions of product types [8, 21, 55]. Alibaba also develops a series of KGs including AliCG [53], AliCoCo [30] and AliCoCo2 [29], where the former two focus on concepts and their IsA relations while the latter one incorporates richer factual relations among concepts. As we have stated in the introduction, there is still a gap between collecting factual knowledge about products and modeling users’ purchasing intentions. Thus, FolkScope is different from the existing product-related knowledge graphs to explicitly and directly model users’ intentions.

Language Models as Knowledge Bases. Large language models, such as BERT [23], GPT [5, 33], and T5 [34], have shown great representation power to revolutionize many downstream applications. It has been shown that language models can memorize factual knowledge, and one can design appropriate prompts to probe knowledge from them, which is usually referred to as “language models as knowledge bases” [1, 32]. Prompts can be either designed by humans, mined from corpus, or tuned from continuous embeddings to improve the knowledge probing tasks [27]. Language models as knowledge bases is a promising idea. However, in practice, we still need symbolic knowledge representation to rapidly support knowledge editing and complicated reasoning. For online recommendations, we sometimes need explicit reasoning to explain the recommendation results. It has been shown that one can derive factual KGs at scale based on language models with proper prompts [12, 43]. For commonsense knowledge, it has been shown that distilling from GPT-3 based on in-context learning with a few examples can generate even human-level commonsense understanding in the form of ATOMIC [44]. None of the above KGs are product-related nor purchasing intention related. Thus, our contribution is unique to the community.

3 KNOWLEDGE GRAPH CONSTRUCTION

In this section, we introduce the knowledge graph construction method using our FolkScope framework.

3.1 FolkScope Framework

We call our framework FolkScope as we are the first attempt to reveal the structure of e-commerce intentional commonsense to rationalize the purchasing behaviors. As shown in Figure 2, FolkScope is a human-in-the-loop approach to semi-automatic construction of the KG. We first leverage the pretrained language models to generate candidate assertions of intentions for purchasing or co-purchasing behaviors based on co-buy data from the Amazon released benchmark dataset. Then we employ two-step annotations to annotate the plausibility and typicality of the generated intentions, where the corresponding definitions of the scores are as follows.

- Plausibility is defined as how possible the assertion is valid regarding their properties, usages, functions, etc.
- Typicality is defined as how well the assertion reflects a specific feature that causes the user behavior. Typical intentional assertions should satisfy the following criteria. Informativeness: The assertion contains key information about the shopping context rather than a general one, e.g., “they are used for Halloween parties”. v.s. “they are used for the same purpose.” Causality: The assertion captures the typical intention of user behaviors, e.g., “they have a property of water resistance.” Some specific attributes or features might largely affect the users’ purchase decisions.
Typicality is different from plausibility as it indicates more typical reasons for purchasing behavior. For example, as shown in Figure 1, “telling the time” is a typical reason for buying a normal watch than buying an “iWatch” whereas a more typical reason may be “fan of Apple products” for the latter one.

After the annotation, we design classifiers to populate the scores to all generated candidates. Then the high-quality ones will be further structured by using pattern mining on their dependency parses to aggregate similar assertions. Then, we also perform conceptualization [51] to further aggregate assertions to form more abstract intentions. Examples are also shown in Figure 1.

### 3.2 User Behavior Data Sampling

We extract the users’ behavior datasets from open-sourced Amazon Review Data (2018) [31] with 15.5M items from Amazon.com. In our work, we mainly consider co-buy pairs, which might indicate stronger shopping intent signals than co-view pairs. After the preprocessing and removing duplicated items, the resulting co-buy graph covers 3.5M nodes and 31.4M edges. The items are organized into 25 top-level categories from the Amazon website, and among them, we choose two frequent categories: “Clothing & Jewelry” and “Electronics” to sample co-buy pairs because those items substantially appear in situations requiring commonsense knowledge to understand while other categories such as “Movie” or “Music” are more relevant to factual knowledge between entities.

Table 1: Statistics of sampled co-buy pairs and generated candidate assertions. Note that the prompts in the generation are not included in the calculations of assertion lengths.

| Type         | Clothing | Electronics | Total |
|--------------|----------|-------------|-------|
| # Item Pairs | 199,560  | 93,889      | 293,449 |
| # Unique Items | 151,509  | 64,244      | 211,349 |
| # Generated Assertions | 11,358,637 | 5,282,273 | 16,640,910 |
| # Unique Assertions | 2,865,118  | 1,280,259 | 4,063,764 |
| Avg. # Tokens of Assertions | 6.66 | 5.25 | 6.21 |

### 3.3 Knowledge Generation

As shown in Table 2, we verbalize the prompt templates using the titles of co-buy pairs. Besides the general prompt (i.e., “open”), we also align our prompts with 18 relations in ConceptNet that are highly related to commonsense. For example, for the relation HasA, we can design a prompt “A user bought item 1 and item 2 because they both have [GEN]” where [GEN] is a special token indicating generation. Since the long item titles might contain noise besides useful attributes, we use heuristic rules to filter out items whose titles potentially affect the conditional generation like repeated words. We use the OPT model [54] of 30B parameters with two NVIDIA A100 GPUs using the HuggingFace library [45] to generate assertion candidates. For each relation of the co-buy pairs, we set the max generation length as 100 and generate 3 assertions using nucleus sampling (p = 0.9) [17]. We post-process the candidates as follows. (1) We discard the generations without one complete sentence. (2) We use the sentence segmenter from Spacy library [3] to extract the first sentence for longer generations. After removing duplicates, we obtain 16.64M candidate assertions for 293K item pairs and 4.06M unique tails among them. The statistics of the two categories are listed in Table 1.

### 3.4 Two-step Annotation and Population

As the generated candidates can be noisy or not rational, we apply the human annotation to obtain high-quality assertions and then populate the generated assertions. We use the Amazon Mechanical Turk (MTurk) platform to annotate our data. Annotators are provided with a pair of co-buy items with each item’s title, category, shopping URL, and three images from our sampled metadata. Assertions with different relations are presented in the natural language form by using the prompts presented in Table 2. To guarantee the quality of our annotation, we set up strict qualification rounds for both plausibility and typicality annotations. Moreover, qualified annotators are provided timely feedback during the annotation.

| Type | Relation | Prompt |
|------|----------|--------|
| Open | HasA     | they both have |
| Item | HasProperty | they both both have a property of |
| Item | RelatedTo | they both are related to |
| Item | SimilarTo | they both are similar to |
| Item | PartOf   | they both are a part of |
| Item | IsA      | they both are a type of |
| Item | MadeOf   | they both are made of |
| Item | CreatedBy | they are created by |
| Item | DistinctFrom | they are distinct from |
| Item | DerivedFrom | they are derived from |
| Function | UsedFor | they both are used for |
| Function | CapableOf | they both are capable of |
| Function | SymbolOf | they both are symbols of |
| Function | MannerOf | they both are a manner of |
| Function | DefinedAs | they both are defined as |
| Human | Result   | as a result, the person |
| Human | Cause    | the person wants to |
| Human | CauseDesire | the person wants his |

---

1. [https://nijianmo.github.io/amazon/](https://nijianmo.github.io/amazon/)
2. [https://huggingface.co/facebook/opt-30b](https://huggingface.co/facebook/opt-30b)
3. As we will further annotate the plausibility and typicality of candidates, larger models will reduce the annotation cost. However, the generation is also constrained by the API cost or the computational cost. Thus, we choose the best model we can use.
4. [https://spacy.io/](https://spacy.io/)
The overall IAA score is 75.48% in terms of pairwise agreement proportion, while the Fleiss’s Kappa [10] is 0.4872.

Different from the simple binary plausibility judgements, in the second step, we have more fine-grained and precise typicality indicators concerning informativeness and causality. Here we choose the candidates automatically labeled as plausible based on our classifier trained on the data in the first step. We ask the annotators to judge whether they are strongly acceptable (+1), weakly acceptable (0.5), rejected (0), or implausible (-1) that the assertion is informative and casual for a purchasing behavior. Thus, if in this step implausible assertion is observed again, they will further decrease the average score. Considering the judgements might be subjective and biased with respect to different annotators, we collect five annotations for each assertion and take the average as the final typicality score. Similar to the first-step annotation, we collect around 60K assertions. Empirically, we find that annotating more candidates does not bring significantly better filtering accuracy. Detailed statistics are presented in Table 3.

### 3.4.2 Population

For plausibility population, we train our binary classifier based on the majority voting results in the first step, which produces binary labels. For the typicality score, as we take the average of five annotators as the score, we empirically use scores greater than 0.8 to denote positive examples and less than 0.2 as negative examples. We split the train/dev sets at the ratio of 80%/20% for both scores and train binary classifiers using both DeBERTa-large [15] and RoBERTa-large [28]. The best trained models are selected to maximize the F1 scores on the validation sets. The results are shown in Table 4. DeBERTa-large achieves better performance than RoBERTa-large on both plausibility and typicality evaluation.

We populate the inference over the whole generated corpus in Table 1 and only keep the assertions whose predicted plausibility scores are above 0.5 (discarding 32.5% generations and reducing from 16.64M to 11.24M). Note that only plausible assertions are kept in the final knowledge graph. Using different confidence cutting-off thresholds leads to trade-offs between the accuracy of generation and the size of the corpus. After the two-step populations, we obtain the plausibility score and typicality score for each assertion. Due to the measurement of different aspects of knowledge, we observe low correlations between the two types of scores (Spearman correlation $\rho$: 0.319 for clothing and 0.309 for electronics).

### 3.5 Knowledge Aggregation

To acquire a knowledge graph with topology structures instead of sparse triplets, we aggregate possible similar assertions. This is done by (1) pattern mining to align similar generated patterns and (2) conceptualization to produce more abstract knowledge.

The assertions are usually expressed as free textual phrases. Some of the phrases are similar that can be merged together by dropping some non-important words, e.g., “they both are skiing products” and “they both are some skiing products.” Therefore, we apply the frequent graph substructure mining algorithm over dependency parse trees to discover the linguistic patterns. We sample 90,000 candidates for each relation to analyze patterns and then parse each candidate into a dependency tree. In addition, the lemmatized tokens, pos-tags, and named entities are acquired for further use. To reduce the time complexity of pattern mining, we mine high-frequency patterns for each relation. To meet the two requirements of the knowledge with high precision but non-trivial, patterns are required to perfectly match more than 500 times. One perfect match means that this pattern is the longest pattern, and no other candidate patterns can match. Therefore, the pattern mining pipeline consists of three passes: (1) a graph pattern mining algorithm, Java implementation of gSpan [47], to mine all candidate patterns with the frequency more than 500, (2) a subgraph isomorphism algorithm, C++ implementation of VF2 algorithm in igraph, with a longest-first greedy strategy to check the perfect match frequency, and (3) human evaluation and revision. Finally, we obtain 256 patterns that cover 80.77% generated candidates.

After pattern mining, we can formally construct our knowledge graph, where the head is a pair of items, the relation is one of the relations shown in Table 2, and the tail is an aggregated assertion that is originally generated and then mapped to a particular one among 256 patterns. Each of the assertions (head, relation, tail) is associated with two populated scores, i.e., plausibility and typicality.

To produce abstract knowledge generalizable to new shopping contexts, we also consider the conceptualization with Probase [14, 46]. This process has been employed and evaluated by ASER 2.0 [51]. The conceptualization process maps one extracted assertion to multiple candidate conceptualized assertions. Finally, we obtain a KG with 184,146 items, 217,108 intentions, 857,972 abstract intentions, and 12,755,525 edges to explain 236,739 co-buy behaviors, where 2,298,011 edges from the view of original assertions and 9,297,500 edges from the angle of conceptualized ones, and 1,160,014 edges model the probabilities of the conceptualization.

### 4 INTRINSIC EVALUATIONS

In this section, we first present some examples of our constructed KG and then conduct comprehensive intrinsic evaluations over KG.

---

3 The annotators in this step are chosen from the high-quality annotators in the first step. We tried other options, such as using seven or nine annotators per generation in our pilot study. The results do not show much improvement.

4 https://github.com/timtadh/parsmis

5 https://igraph.org/
Table 5: Two examples from the constructed knowledge graph. “P.” and “T.” mean the model predicted plausibility and typicality scores respectively. Generated tails with high typicality (in green) and low typicality (in red) scores are highlighted.

| Item 1 | Item 2 | Relation | Tail |
|--------|--------|----------|------|
| GGS III LCD Screen Protector glass for CANON 5D Mark III (link) | ECCSD3B Secure Grip Camera Case for Canon 5D Mark III (link) | Open | they can be used for the same purpose |
| | | HasProperty | "easy to install" and "easy to remove" |
| | | SimilarTo | the product he bought |
| | | PartOf | his camera gear |
| | | UsedFor | protect the camera from scratches and dust |
| | | SymbolOf | his love for his camera |
| | | DefinedAs | "Camera Accessories" on Amazon.com |
| Sun Smarties Baby Girls UPF 50+ Non-Skid Sand and Water Socks Small Hot Pink (link) | Schylling UV Play Shade, SPF 50+, Ultra portable , Blue (link) | Open | he was worried about his baby’s skin |
| | | SimilarTo | each other |
| | | DistinctFrom | other similar products |
| | | UsedFor | baby’s outdoor activities |
| | | CapableOf | blocking harmful UV rays |
| | | DefinedAs | sun protection products |
| | | Result | enjoy the sun sagely and comfortably |
| | | Cause | want to use them for his/her baby |

Table 6: Acceptance ratios of plausible assertions and the corresponding sizes of populated assertions with different cutting-off thresholds.

| Threshold | Clothing | Electronics | Total |
|-----------|----------|-------------|-------|
|           | Accept   | Accept      | Accept | Size   | Size   | Size   |
| 0.5       | 83.73%   | 82.74%      | 83.40% | 3,250,605 | 11,236,636 |
| 0.7       | 90.27%   | 88.27%      | 90.40% | 10,214,416 |
| 0.8       | 91.02%   | 89.50%      | 90.00% | 9,598,231 |
| 0.9       | 95.60%   | 94.87%      | 95.36% | 8,397,738 |

4.1 Examples

We show two examples of co-purchasing products and their corresponding knowledge (§ 3.3) as well as populated scores (§ 3.4) in Table 5. We measure the quality of generated quality using both plausibility and typicality scores, which are again shown they are not correlated. For example, “they are SimilarTo the product they bought” for the first pair and “they are DistinctFrom other similar products” for the second pair are plausible assertions but not typical explanations of why a user would buy them together. Moreover, some of the open relations are very good as well. Take the second pair as an example: the open relation shows "he was worried about his baby’s skin" as both products are related to baby skin protection.

4.2 Human Evaluation

As we populate the whole generated assertions using two trained classifiers based on DeBERTa-large model and thus all of assertions have been labeled with the plausibility and typicality scores. To further evaluate the effectiveness of the knowledge population, we conducted the human evaluations by sampling a small number of populated assertions from different scales of predicted scores.

4.2.1 Plausibility Evaluation. We randomly sample 200 plausible assertions from each relation in each of the clothing and electronics domains to test the human acceptance rate. The annotation is conducted in the same way as the construction step. As we only annotate assertions predicted to be greater than the 0.5 plausibility score, the IAA is above 85%, even greater than the one in the construction step. As shown in Table 6, different cutting-off thresholds (based on the plausibility prediction by our model) lead to the trade-offs between the accuracy and the KG size. Overall, FolkScope can achieve 83.4% acceptance rate with a default threshold (0.5). To understand what is filtered, we manually check the generations with low plausibility scores and find that OPT can generate awkward assertions, such as simply repeating the item titles.
or obviously logical errors regarding corresponding relations. Our classifier trained on annotated datasets helps resolve such cases. Using a larger threshold 0.9, we attain 95.35% acceptance rate, a nearly 11.96% improvement while still keeping above 8M plausible assertions. We also report the accuracy in terms of different relations in Table 7. We can observe that assertions concerning the relations of human beings’ situations like Cause, Result, and CauseDesire have relatively lower plausibility scores and longer lengths than the relations of items’ property, function, etc. This is because there exist some clues about items’ knowledge in the item titles, while it is much harder to generate (or guess) implicit human-being’s casual reasons using language generation.

4.2.2 Typicality Evaluation. The goal of the typicality population is to precisely recognize high-quality knowledge, and we evaluate whether assertions with high typicality scores are truly good ones. We randomly sample 200 assertions from each relation whose predicted typicality scores are above 0.8 for human evaluation. Each item titles, while it is much harder to generate (or guess) implicit human-beings’ casual reasons using language generation.

| Threshold | Aggregated Knowledge | Conceptualization |
|-----------|----------------------|-------------------|
| 0.8       | 62.15%               | 45.71%            |
| 0.9       | 63.35%               | 55.67%            |
| 0.99      | 70.28%               | 57.75%            |

Figure 3: Average typicality score of each relation in the populated KG with the cutting-off threshold 0.8.

Table 9: The generated knowledge in the same subcategory.

| Subcategory | Generation |
|-------------|------------|
| (Costumes, Toys) | he wants to disguise himself as a superhero, he wanted to be a star war character for Halloween, they are both a manner of Christmas decoration, he wants kids to have fun and enjoy the Easter holiday, he is able to dress up as a pirate |
| (Dresses, Dresses) | they are symbol of the fashion trend, they can both be worn for casual occasions, they are both used for wedding dress, they are both capable of giving a good fit, they can both being worn by girls of any age |

Table 9: The generated knowledge in the same subcategory.

HasPropertyOf are less typical compared to other relations. They describe items’ general features but can not well capture typical purchasing intentions though they have high plausibility scores, whereas CapableOf and MadeOf are the most typical features that can explain purchasing intentions for the two domains we concern.

4.3 Novelty Evaluation

As we know, language model based generation capture spurious correlation given the condition of the generation [20]. Hence we simply quantify the diversity as the novelty ratio of generated tails not appearing in the item titles, i.e., novel generations. For example, the title “Diesel Analog Three-Hand - Black and Gold Women’s watch” contains specific attributes like “Black and Gold” or type information “women’s watch.” Such knowledge can be easily extracted by off-the-shelf tools. Traditional information extraction based approaches mostly cover our knowledge if the generation simply copies titles to reflect the attributes. Otherwise, it means that we provide much novel and diverse information compared with traditional approaches. The novelty ratio increases from 96.85% to 97.38% after we use the trained classifiers for filtering. Intuitively, filtering can improve the novelty ratio. For the assertions whose typicality scores are above 0.9, we also observe that the novelty ratio reaches 98.01%. These findings suggest that FolkScope is indeed an effective framework for mining high-quality novel knowledge.

4.4 Fine-grained Subcategory Knowledge

Since the items are organized in multilevel fine-grained subcategories in the catalog of shopping websites, we are interested in whether our constructed KG contains high-quality common intentions among items belonging to subcategories. The common knowledge can be useful to have intention-level organizations besides category-level and further help downstream tasks. The co-buy item-pairs in our sampled clothing category fall into 15,708 subcategory pairs, such as (necklaces, earning) or (sweater, home & kitchen), where most of them are different subcategories in one pair. We select frequent common assertions with high typicality scores to demonstrate the abstract knowledge. Two examples are shown in Table 9. Though costumes and toys belong to two different types, they are complementary because of the same usage, such as "Halloween," "Easter holiday," and "Christmas," or sharing the same key

| Subcategory | Generation |
|-------------|------------|
| (Costumes, Toys) | he wants to disguise himself as a superhero, he wanted to be a star war character for Halloween, they are both a manner of Christmas decoration, he wants kids to have fun and enjoy the Easter holiday, he is able to dress up as a pirate |
| (Dresses, Dresses) | they are symbol of the fashion trend, they can both be worn for casual occasions, they are both used for wedding dress, they are both capable of giving a good fit, they can both being worn by girls of any age |
5 EXTRINSIC EVALUATION

In this section, we evaluate the constructed KG by demonstrating it can improve recommendations. Further analysis shows that the performance can improve with more plausible and typical edges.

5.1 Experimental Setup

To keep consistent with the constructed KG, we also use user-item interaction datasets of the two categories from the Amazon Review dataset [31]. The detailed statistics are in Table 10. Each dataset is separated into training, validation, and testing sets with a ratio of 8:1:1. We use the root mean square error (RMSE) to evaluate the performance of the recommendation system. All experiments repeat five times, and the average scores are reported.

To fairly evaluate the KG for recommendations, we restrict the usage of the KG such that we can only use the co-buy pairs that are simultaneously purchased by at least one user in the training set. With this restriction, we match a sub-graph of the original KG by using the actual co-buy from the users in the recommendation training set. The detailed statistics of the matched KG are in the first line of Table 11. The item coverage computes the percentage of the items in the recommendation dataset that are covered by the matched KG. We also want to evaluate whether two rounds of annotations and populations on plausibility and typicality can improve the recommendation results. We further filter the matched KG with a threshold of 0.5 or 0.9 on both two scores. In Table 11, the number of edges essentially reduces when the filters are applied, but the coverage of the items does not drastically drop.

To incorporate the KG information into the recommendation task, we first learn item embeddings from the matched KG, and then use the embeddings as features for the Wide&Deep model [6] to train a recommendation model. To learn the item representations from a matched KG, we use a modified version of the TransE [3] algorithm. The modifications are as follows. First, we initialize all the node embeddings of the tail nodes with their Sentence-BERT [37] representations. Second, as the head of a triple is a pair of items, we compute the average of item embeddings as the head embedding for the TransE model. In the negative sampling process, either one or two items in the head are randomly corrupted. The objective function is then described by the following equation:

$$\mathcal{L} = y \cdot d\left(\frac{p_1 + p_2}{2} + r, e\right) - d\left(\frac{p_1' + p_1''}{2} + r, e\right),$$

where $y$ is a margin parameter, and $p_1, p_2, p_1', p_1''$ are the positive and negative item embeddings for items $p_1, p_2, p_1', p_1''$, respectively. Meanwhile, $r$ is the relation embedding for relation $r$, $e$ is the embedding for the tail $e$, and the function $d$ is Euclidean distance.

5.2 Experimental Results

We conduct two ablation studies to evaluate the effect of structural information provided by the co-buy pairs and the semantic information provided by the tails’ text only. For the former, we train a standard TransE model solely on co-buy pairs to learn the graph embeddings of items. For the latter, for each item in the matched KG, we conduct average pooling of its neighbor tails’ Sentence-BERT embeddings as its semantic representations. The experimental results are shown in Table 12, and we have the following observations. First, the textual information contained in intentional assertions is useful for product recommendations. This can be testified as the W&D model can perform better even when only features of the assertions are provided. Second, our KG, even before annotations and filtering, can produce better item embeddings than solely using the co-buy item graphs. As we can see, the performance of our matched KG is better than that of the co-buy pair graphs. Third, the two steps of annotation and population indeed help improve the item embeddings for recommendations. The higher the scores are, the larger improvement the recommendation system obtains.

| Method                | Clothing | Electronics |
|-----------------------|----------|-------------|
| NCF [16]              | 1.117    | 1.086       |
| W&D [6]               | 1.104    | 1.071       |
| + Co-Buy Pairs Features Only | 1.096    | 1.067       |
| + Eventuality Features Only | 1.093    | 1.068       |
| + Matched Knowledge Graph | 1.093    | 1.058       |
| + Plausibility > 0.5  | 1.087    | 1.060       |
| + Plausibility > 0.5 and Typicality > 0.5 | 1.081    | 1.053       |
| + Plausibility > 0.9  | 1.086    | 1.053       |
| + Plausibility > 0.9 and Typicality > 0.9 | 1.081    | 1.052       |

6 CONCLUSION

In this paper, we propose a new framework, FolkScope, to capture intentional commonsense knowledge for e-commerce behaviors. We develop a human-in-the-loop semi-automatic way to construct an intention KG, where the candidate assertions are automatically generated from large pretrained language models, with carefully designed prompts to align with ConceptNet commonsense relations.
REFERENCES

[1] Badr AlKhamissi, Millicent Li, Adi Celikyilmaz, Mona Diab, and Marjan Ghazvininejad. 2022. A review on language models as knowledge bases. arXiv:2204.06031 (2022).

[2] Kurt D. Bollacker, Colin Evans, Praveen K. Paritosh, Tim Sturge, and Jamie Taylor. 2004. Freebase: A collaboratively created graph database for structuring human knowledge. In SIGMOD. 1247–1256.

[3] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. Advances in neural information processing systems 26 (2013).

[4] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chattaniya Malaviya, Adi Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. In the ACL. 4762–4779.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. NeurIPS (2020), 1877–1901.

[6] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Colin Raffel, et al. 2020. Flan-T5: Pushing the limits of zero-shot text-to-text transfer. In NAACL-HLT. 1, 8 (2019), 1981–1991.

[7] Qifan Cai, Daniel Hutto, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. 21, 140 (2020), 1–67.

[8] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In EMNLP. 2383–2392.

[9] Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. 2018. EventMind: Commonsense Inference on Events, Intents, and Reactions. In ACL. 463–473.

[10] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. 721–724.

[11] Shubho Hao, Bowen Tan, Kaiwei Tang, Hengzhong Zhang, Eric P. Xiong, and Zhiting Hu. 2022. BertNet: Harvesting Knowledge Graphs from Pretrained Language Models. arXiv:2206.14282 (2022).

[12] Zhenyun Hao, Jianing Hao, Zhaozhao Peng, Senczhang Wang, Philip S Y, Yue Wang, and Jian Li. 2022. Ky-HEIN: Dynamic Evolution based Deep Hierarchical Empirical Intention Network for Membership Prediction. In WSDM. 363–371.

[13] Mutian He, Tianqing Zhang, Feiqing Wang, and Yangqiu Song. 2022. Acquiring and Modelling Abstract Commonsense Knowledge via Conceptualization. arXiv preprint arXiv:2206.01532 (2022).

[14] Pengcheng He, Xinyi Cai, Jinliang Gao, and Weizhu Chen. 2021. DeBERT: Decoding-enhanced BERT with Disentangled Attention. In ICLR.

[15] Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying Recommendations for Recommendation. In SIGKDD. 217–218.

[16] Ari Holtzman, Jan Buys, Li, Rui Ge, and Yejin Choi. 2019. The Curious Case of Natural Text Degeneration. In ICLR.

[17] Daniel Hutto and Ian Ravenscroft. 2012. Folk Psychology as a Theory. In Metaphysics Research Lab, Stanford University.

[18] Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff De, Kirouke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (Comet-) atomac: 2020: on symbolic and neural commonsense knowledge graphs. In AAAI 6384–6392.

[19] Ziwei Ji, Nayeon Lee, Rita Frieske, Tizheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in neural commonsense knowledge graphs. In Workshop on Collaboratively-Built Knowledge Sources and Artificial Intelligence.

[20] Yoshio Iwata, Weiwei Wu, Jinlan Ju, Zhenlong Yang, and Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. arXiv:2107.15836 (2021).

[21] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. arXiv:1907.11692 (2019).

[22] Xusheng Luo, Le Bo, Jinjung Wu, Lin Li, Zhiy Luo, Yonghua Yang, and Keping Yang. 2021. AlCoCo2: Commonsense Knowledge Extraction, Representation and Application in E-commerce. In SIGKDD. 3385–3393.

[23] Xusheng Luo, Luxin Liu, Yonghua Yang, Le Bo, Yuanpeng Cao, Jinghua Wang, Qiang Li, Keping Yang, and Kenny Q. Zhu. 2020. AlCoCo: Alibaba e-commerce cognitive concept net. In SIGMOD. 313–327.

[24] Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. In EMNLP. 188–197.

[25] Fabio Petrotoni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Müller. 2019. Language Models as Knowledge Bases?. In IJCAI. 3638–3645.

[26] Nils Reimers and Iryna Gurevych. 2020. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing.

[27] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for d:then reasoning. In the AAAI. 3027–3035.

[28] Yangqiu Song, Haixun Wang, Zhongyuan Wang, Hongming Zhang, and Weizhu Chen. 2011. Short Text Conceptualization Using a Probabilistic Knowledgebase. In IJCAI. 2330–2336.

[29] Rohin Shah,(resp) and the AI collaboration.

[30] Yi Feng and Jiawei Han. 2002. gSpan: Graph-Based Substructure Pattern Mining. In ICDM. 721–724.

[31] Nassir Zalmout, Chenwei Zhang, Xiaoli Li, Yan Liang, and Xin Luo. 2021. All You Need to Know to Build a Product Knowledge Graph. In SIGKDD. 4090–4091.

[32] Chenwei Zhang, Wei Fan, Nan Du, and Philip S Yu. 2016. Mining user intentions for Recommendation. In SIGKDD. 755–764.

[33] Hongming Zhang, Daniel Khashabi, Yangqiu Song, and Dan Roth. 2020. Tran–OMS: From Linguistic Graphs to Commonsense Knowledge. In IJCAI. 4044–4050.

[34] Hongming Zhang, Xin Liu, Haojie Pan, Haowen Ke, Jiefu Du, Tianqing Fang, and Yangqiu Song. 2022. ASER: Towards large-scale commonsense knowledge acquisition via higher-order selectional preference over eventualities. Artif. Intell. 309 (2022), 103740.
Survey Instructions (Click to Collapse)

E-Commerce Behavior Assertion Validation

Hi there! Welcome to our HITs! In this task, you are required to select the most plausible assertion(s) for a prompted e-commerce behavior and provide brief reasons for your decisions.

You are advised to grab information from the Amazon link page if there's an assertion that you are not sure about. If the link doesn't work, please try to google for some additional information related to both items. Please stick to the content of each assertion.

You are being paid $1 per HIT. You can complete the task in any order. You will receive an average of 1 HIT per minute. A maximum of 60 minutes is allotted for this HIT. You can stop at any time. Please answer all questions, even if you are not sure about the answers.

This message is only visible to you and will not be shown to Workers.

Previewing Answers Submitted by Workers

[52] Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. 2020. ASER: A large-scale eventuality knowledge graph. In WWW. 201–211.
[53] Nangyi Zhang, Qiaonghai Ju, Shumin Deng, Xiang Chen, Hongbin Ye, Hai Chen, Huaxiao Tou, Gang Huang, Zhao Wang, Nengwei Hua, et al. 2021. Alcg: Fine-grained and evolvable conceptual graph construction for semantic search at alibaba. In SIGKDD. 3985–3995.
[54] Sunan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Muya Chen, Shouchen Chen, Christopher Dewan, Mona Dab, Xian Li, Xi Victoria Lin, et al. 2022. Opt-Open pre-trained transformer language models. arXiv:2205.01068 (2022).
[55] Xinyang Zhang, Chenwei Zhang, Xian Li, Xin Luna Dong, Jingbo Shang, Christos Faloutsos, and Jianwei Han. 2022. OA-Mine: Open-World Attribute Mining for E-Commerce Products with Weak Supervision. In WWW. 3153–3164.

APPENDIX

Annotation Guideline

Workers satisfying the following three requirements are invited to participate: (1) at least 90% lifelong HITs approval rate, (2) at least 1,000 HITs approved, and (3) achieving 80% accuracy on at least 10 qualification questions, which are carefully selected by authors of this paper. Qualified workers will be further invited to annotate 16 tricky assertions. Based on workers’ annotations, they will receive personalized feedback containing explanations of the errors they made along with advice to improve their annotation accuracy. Workers surpassing these two rounds are deemed qualified for main round annotations. To avoid spamming, experts will provide feedback for all workers based on a sample of their main rounds’ annotations from time to time.

We conduct human annotations and evaluations on the Amazon Mechanical Turk using the Figure 4 for the first-stage plausibility annotation and the Figure 5 for the second-stage typicality annotation.

Knowledge Population

Using different confidence cutting-off thresholds leads to trade-offs between the accuracy of generation and the size of the corpus. Higher values result in conservative selections that favor precision over recall, whereas lower ones tend to recall more plausible assertions. We plotted four cutoff points in Figure 6.

Figure 4: The question card in our plausibility annotation round. A prompted assertion with its corresponding relation is presented to Turk worker. Workers can choose one from valid, invalid, and unfamiliar.

Figure 5: The question card in our typicality annotation round. A prompted assertion with its corresponding relation is presented to Turk worker. Workers can choose one from valid, invalid, and unfamiliar.

Figure 6: The precision-recall curve of our plausibility population classifier on the human-labeled validation set. The annotated points show the different thresholds (cutoffs) to filter the generated assertions, i.e. from left to right: 0.9, 0.8, 0.7, 0.5 respectively.

Table 13: Frequent pattern coverage on human-annotated knowledge.