Building Knowledge Graphs of Experientially Related Concepts

Wenjie Yang WYANGBC@CONNECT.UST.HK
Xiaojuan Ma MXJ@CSE.UST.HK
Department of Computer Science and Engineering,
Hong Kong University of Science and Technology, Hong Kong, China

Abstract

Consumers assess and select products and services based on a combination of objective factual attributes (e.g., price) and subjective experiential factors. For example, when choosing a restaurant, users often focus on the food quality and ambiance. State-of-the-art search services provide powerful interfaces for filtering objective properties but struggle to support users through the process of considering experiential factors. One of the key reasons for this discrepancy is that the objective properties are clearly represented by a database schema, but there is no such equivalent for experiential properties, which are vaguer by nature. This paper introduces CoNex, a pipeline for building knowledge graphs (KGs) that describe concepts concerning consumers' experiences in a given domain and the relationships between them. CoNex begins by harvesting experience-related concepts on a domain-specific corpus and then discovering experiential connections between them. CoNex further expands its knowledge coverage by a pre-trained language model fine-tuned via data from hybrid sources. Our experiments demonstrate that the KGs constructed by CoNex accurately reflect the experiential relationships between concepts as judged by humans. Finally, we show the effectiveness of using these KGs as tools to improve the performance of an experience-oriented search task.

1. Introduction

Consumers are known to use objective and functional features to reason their judgments and choices of products and services [Shafir et al., 1993, Simonson, 1989]. However, they also evaluate products based on subjective and experiential factors such as restaurant ambiance and hotel cleanliness [Nelson, 1970, Alba et al., 1997, Huang et al., 2009]. Prior studies have looked into how experiences arise in examining and consuming a product (e.g., [Arnould, 2004]) and how to create compelling consumer experiences [Brakus et al., 2014]. However, without a comprehensive picture of the diverse factors that may contribute to the overall experience of a product or service, it can be challenging to support experience-related product/service design, search (e.g., [Evensen et al., 2019, Li et al., 2019]), discovery, and personalization [Dong, 2018]. Many works have attempted to construct databases and knowledge graphs (KGs) of the functional product features (e.g., [Xu et al., 2020]), but there still lacks research that systematically charts consumer experiences in a given domain.

This work aims to fill this gap and frames the problem as building KGs that represent consumer experiences in given domains with experientially connected concepts. We propose to tackle this problem by extracting experience-related knowledge and inferring its relationality from semi-structured, domain-related sources (e.g., text reviews) and from structured

corpora (e.g., commonsense knowledge bases). We call our resulting KGs *experience knowledge graphs*. The vertices in an experience KG are *concepts* that people experience when consuming the target product/service, and a directed edge between two vertices denotes *experiential relatedness*. In other words, "$A \rightarrow B$" indicates that the experience with a *concept* A is affected by the experience with another *concept* B, which is usually more specific than A. For example, when choosing a hotel, people commonly focus on its service and cleanliness[Li et al., 2019]; that is, the experience of both factors affects the experience of the hotel. This can be represented by "`hotel` $\rightarrow$ `service`" and "`hotel` $\rightarrow$ `cleanliness`" in the resulting experience KG.

The most accurate way of acquiring the relations between consumer experience factors is through expert design [Li et al., 2019] or crowdsourcing, but these methods are rather expensive and difficult to scale up and apply to new domains. An alternative approach is to use human language and textual expressions as proxies and employ computational methods to mine such relatedness from them. Earlier NLP research has shown that such an approach can effectively reveal factual and semantic relationships [Havasi et al., 2007, Wu et al., 2012, Gabrilovich et al., 2007, Radinsky et al., 2011] between concepts. However, these relationships are typically generic and may not reflect the connections between factors concerning people's subjective experiences in a specific domain.

Mining experiential relatedness between concepts from text as a novel task faces several challenges. 1) Experience-oriented relationships are usually not explicitly expressed in natural languages; even if they are, it is difficult to capture such relatedness in fixed syntactic patterns due to the diversity of expressions [Halevy, 2019]. 2) Language usually only reveals the most salient connections between concepts, as many associations are internalized and rarely expressed in language [De Deyne et al., 2016]. This might limit the coverage of experiential associations extracted by corpus-based approaches from a single source.

To address these challenges, we propose CoNex, a pipeline for automatically mining concepts and relationships in a KG of consumer experiences by leveraging both domain and external knowledge. CoNex includes two steps: 1) build a basic experience KG based on domain-specific user reviews by identifying experience-related concepts through an opinion extractor and inferring their relatedness based on distributional similarity. 2) extend the basic KG's coverage with a pre-trained language model, BART [Lewis et al., 2020] in particular. It is fine-tuned on both domain-specific data, including samples from the basic KG and opinion implications [Bhutani et al., 2020], and suitable external data such as word association [De Deyne et al., 2019] or physical commonsense [Speer et al., 2017, Hwang et al., 2020].

We assess the performance and generalizability of CoNex across different domains, including hotels, restaurants, and electronics. We conduct a human evaluation on Amazon Mechanical Turk, and the results suggest that the basic KG constructed by CoNex is more accurate and has better coverage than other baselines in capturing experientially related concepts, and integrating external knowledge into CoNex can further widen the performance gap. We also demonstrate the effectiveness of our KGs (basic and extended) in an experience-oriented IR task based on SubjQA [Bjerva et al., 2020], which simulates search services that find answers to consumers' experience-related questions from textual

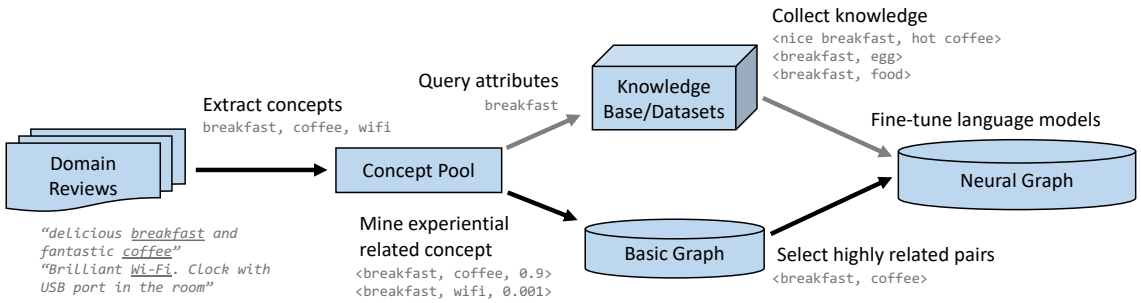


Figure 1: CoNex's pipeline: 1) mine experience-related knowledge from a domain-specific corpus to form a basic graph (indicated by black arrows); 2) learn from domain-specific and external data to generate more comprehensive knowledge from constructed neural graphs (grey arrows).

user reviews. The results show that the CoNex-empowered classifier can achieve better F1 scores than other baselines. We release our code for future research[1].

2. CoNex

2.1 Experience Knowledge Graph Construction Using Domain Corpora

Economists define experiential factors used by consumers to evaluate products and services as those that cannot be evaluated prior to purchasing (or using) products or services [Nelson, 1970, Klein, 1998]. They observe that people tend to refer to other people's reviews when assessing these factors, especially as e-commerce becomes more prevalent [Alba et al., 1997]. In light of this, we may be able to mine experiential factors using the lens of online user reviews.

As shown in Figure 1, the first step of CoNex is to analyze user reviews as a domain-specific corpus, extract concepts that represent consumers' experiences, and use a distributional approach to capture the experiential relatedness between them.

Concept Extraction. CoNex starts by using an opinion extractor to identify as many concepts as possible from the domain corpus. Opinion extractors identify opinions as $<$ *modifier*, *aspect* $>$. We keep only the aspect part as concepts. Opinion extractors typically include deep learning-based extractors and rule-based extractors, which may differ in their performance like recall. CoNex does not rely on the type of extractor. In our experiments, we used a state-of-the-art extractor [Li et al., 2019] in the hotel and restaurant domains, while in the electronics domain, we used a rule-based extractor provided by [Bhutani et al., 2020]. All identified concepts form a concept pool.

Computing Experiential Relatedness by Distributional Method. Next, CoNex represents concepts with bag of words (BoW) model and measures their experiential relatedness with cosine similarity. Specifically, for each concept, we collect the concepts that

1. https://github.com/ywj-cs/CoNex

co-occur with it in a context window across all reviews and add them to the vector along with the frequencies. The context window size N limits the number of sentences considered in the collection of co-occurring concepts. In this paper, we only consider $N = 1$, i.e., the concept vector counts only the frequencies of co-occurring concepts in the same sentence. Then, we use cosine similarity to represent the relatedness between concept vectors and construct a nearest neighbor graph with concept as the node and relatedness as the edge. This forms a basic experience KG of concept pairs.

The BoW-based distributional method is already able to identify some of the experiential related concepts. However, the coverage of this approach may be limited by many factors, such as methodological limitations or the fact that some experiential associations are rarely expressed in language. The next step of CoNex is to expand the knowledge covered in the domain from other sources of information.

2.2 Language Models as Neural Knowledge Graphs

Recent work has shown the effectiveness of using language models as neural knowledge graphs to obtain more hypothetical knowledge [Bosselut et al., 2019, Hwang et al., 2020]. Inspired by this work, a straightforward idea is to use language models as another source to uncover more experiential relationships between concepts. However, fine-tuning language models require large and accurate training data. Although we can filter out high-quality parts of the base graph, the amount of data is often insufficient for the models to learn. To address these issues, CoNex collects complementary data from different sources and fine-tunes BART [Lewis et al., 2020] — a state-of-the-art denoising sequence-to-sequence pre-trained language model, through a knowledge integration framework in order to obtain neural KGs with broader knowledge coverage.

Knowledge Collection. CoNex collects knowledge from both inside and outside the domain with the following relationships: 1) **Opinion implication** [Bhutani et al., 2020] describes the subjective and implicit connection between consumers' opinions in the domain, such as *good meal*s implies *good restaurant*s; 2) **Physical common sense** [Havasi et al., 2007] entails the encyclopedic relationship between objects, such as (*hotel, UsedFor, sleeping*); 3) **Word association** [De Deyne et al., 2019] is a psychological game in which, when presented with a word, participants need to respond with the first word that comes to mind. People would often respond to the word *nurse* when they are presented with the word *doctor*. We remove the relation parts in common-sense data and represent all sources of knowledge in the form of tuples (*head*, *tail*).

CoNex then selects appropriate training data from these sources and the base graph to fine-tune BART. For each concept in the basic KG, we first rank its neighbors (i.e., related concepts) based on cosine similarity and then set the maximum ranking (i.e., top k) and the minimum similarity. Those neighbors that do not satisfy either condition are removed, resulting in a set of highly related concept pairs. Then we retrieve the data with the same head as these pairs in the above three sources.

Learning from Hybrid Knowledge. We developed a knowledge integration framework to integrate these data from different sources, training BART to generate multi-source knowledge simultaneously. The head of the collected data is prefixed with its source (e.g.,

concept and *common sense*) and then used as input to BART, while the tail of the data serves as the training target. Then, the size of the data from different sources is balanced using example-proportional mixing [Raffel et al., 2020]. Two kinds of domain data (i.e., experiential related concepts and opinion implications) are included in BART's training set. Note that we only add one type of external knowledge to the training set at a time in order to investigate how different types of knowledge affect the model's performance in different domains. Nonetheless, we did attempt to train the model with both types of knowledge, but it performed no better than with either. This could be due to the fact that introducing too much external knowledge introduces noise, causing the model to fail to converge.

After fine-tuning the training data, the model can be used to generate domain-specific related concepts with the specified prefix (i.e., *concept*).

3. Experiment

We perform two kinds of experiments across three domains to evaluate different experience KGs constructed by CoNex and baselines: 1) we first use human-based evaluation to examine the quality of the experientially related concept pairs captured by these KGs, and then 2) we created an answer retrieval task based on SubjQA [Bjerva et al., 2020] to investigate the efficacy of using these KGs in downstream applications, especially experience-oriented search.

Domain	# Reviews	# Concepts	# Training set*			Example Concepts
Hotel	1.55M	216K	10K	103K	87K/80K	`room, service`
Restaurant	1.44M	525K	10K	81K	79K/89K	`food, drink`
Electronics	8,74M	822K	4K	71K	63K/61K	`screen, sound`

Table 1: Number of reviews, extracted concepts, and the training set per domain. *include pairs of concepts, opinion implications, word association, or common senses.

Datasets We create experience KGs for three domains, including hotels, restaurants, and electronics, based on the reviews datasets from Tripadvisor [Marcheggiani et al., 2014], Yelp[2], and Amazon [He and McAuley, 2016]. The experience-related concepts are extracted from each review using the opinion extractor described in Section 2.1. Table 1 shows the statistics of the datasets we used and the number of concepts we obtained. We use existing datasets as the source of additional knowledge required in the second step of CoNex, including Small World of Words (SWOW) [De Deyne et al., 2019] for word association and ConceptNet [Speer et al., 2017] and ATOMIC$^{20}_{20}$ [Hwang et al., 2020] for common sense. Following Hwang et al., we remove ConceptNet triples with a weight below 0.5 or a negative relation (e.g. NotIsA) to ensure quality. For ATOMIC$^{20}_{20}$, we retain only triples with a physical relation. We use SAMPO [Bhutani et al., 2020] to obtain knowledge of the

2. https://www.yelp.com/dataset

opinion implication for each domain, in which we remove the opinion pairs with similarity below 0.875[3].

Baselines and Settings The two-step pipeline of CoNex creates different KGs based on data from different sources. We obtain these KGs by ablation and compared their quality in the experiments to show the pipeline's effectiveness.

In the first step of CoNex, we construct a basic graph (called **BoW**) based on the domain dataset. The second step entails filtering high-quality samples from BoW and using them to retrieve external data from three sources. We apply the filtering approach described in Section 2.2, where the maximum ranking k for the hotel domain is 10, k for the other two domains is 30, and the minimum similarity is 0.8 for all domains. The BoW samples are divided into training, validation, and test sets by 8:1:1 and then the training set and the data from three sources are balanced by example-proportional mixing. By gradually adding these data to the training set, we can obtain a series of fine-tuned BARTs as neural KGs, specifically using BoW samples for training to get **K(BoW)** and adding SAMPO data to get **K(SAMPO)**. Then we choose to add SWOW data to get **K(SWOW)** or choose to add common sense to get **K(CM)**. The training details are given in Appendix A.1. The sizes of the training sets containing data from these sources are also presented in Table 1.

We also use **SAMPO** as a baseline for constructing experience KGs. Although SAMPO is designed for building domain-specific KGs of opinions and their implication relations, it can still be used to assess experiential relatedness between concepts. Since they use matrix decomposition to produce the embeddings of opinions $< \textit{modifier}, \textit{aspect} >$, we can treat aspects embeddings as concept embeddings and measure their relatedness by cosine similarity, as we do in CoNex. We use the optimal parameters they report for producing these embeddings.

3.1 Human Evaluation

Six KGs (one created by baseline SAMPO and the rest by CoNex) were first evaluated based on whether their experientially related concept pairs were acceptable to people. Following Bhutani et al.'s evaluation settings, we used weighted random sampling to select a set of concepts from each domain dataset, using concept frequencies as weights and removed any concept that could be considered noise introduced by the opinion extractor, and ultimately kept 200 concepts per domain.

Given each concept, we collected the five most relevant neighbors from every KG to form pairs with the input concept. In BoW (our basic graph) and SAMPO, we select the neighbors with the highest cosine similarity, and in the neural KGs, we use beam search (size = 5) to generate these top neighbors. In each concept pair ($A \rightarrow B$), three Amazon Mechanical Turk workers were asked to judge (accept/reject/not sure) whether the experience of A is affected by the experience of B in the corresponding domain. The majority vote is taken as the label for the pair. In total, N=18,000 judgments were made per domain (200 concepts $\times$ 3 workers $\times$ 5 neighbors $\times$ 6 KGs). A detailed description of

3. Bhutani et al. prune neighbors based on cosine similarity < 0.8; we improve such a threshold to 0.875 in order to maintain a higher quality for BART training.

Metrics	Hotel Model 1	2	3	4	5	6	Restaurant Model 1	2	3	4	5	6	Electronics Model 1	2	3	4	5	6
✓(↑)	50.5	73.5	79.4	79.9	**86.6**	**86.7**	58.2	78.7	64	63.5	**82.5**	79	39.8	64.6	60.4	43.7	64.4	**74.4**
✗(↓)	47.8	24.4	17.8	18.5	13	**12.1**	34.8	17	34.5	34.5	**16.7**	16.8	58.1	32.7	38	54.8	33.7	**23.9**
? (↓)	1.7	2.1	2.8	1.6	**0.4**	1.2	7	4.3	1.5	2	**0.8**	4.2	2.1	2.7	**1.6**	**1.5**	1.9	1.7
MAP (↑)	39.6	66	73.1	72.5	**80.4**	**80.5**	48.5	71.5	52.3	51.4	**75.2**	70.3	29.6	56.1	50.5	32.1	54.3	**65.9**

Table 2: The human evaluation of models 1 to 6, including SAMPO, BoW, K(BoW), K(SAMPO), K(SW), and K(CM), in terms of rates of acceptance ✓, rejection ✗, uncertainty ?, and MAP across three domains.

the MTurk task design is provided in Appendix A.3. Following Bhutani et al.'s work, we then use the annotated set[4] to compute the precision and "pseudo-recall" for each KG.

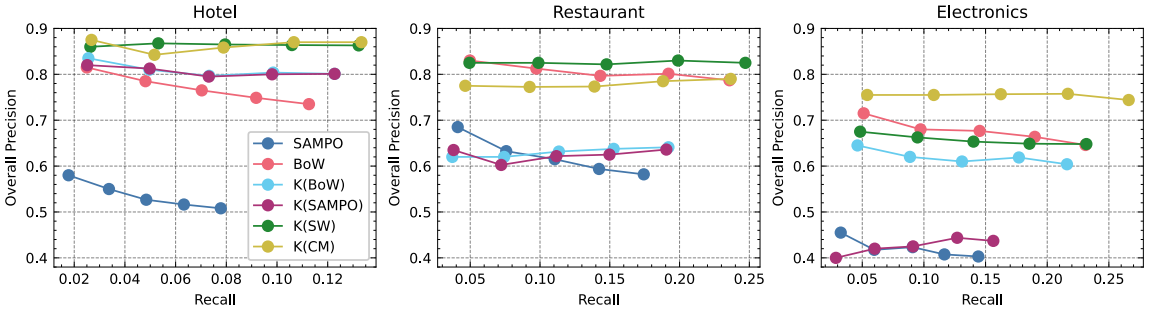


Figure 2: PR curves of models at different values of top k (1-5) across three domains.

Results and Analysis We report the acceptance rate and the mean average precision (MAP) for each graph in Table 2 and compute the overall precision and pseudo-recall for different values of top k (1-5), and present them across a PR curve (Figure 2). The results demonstrate the effectiveness of CoNex's use of the distributional method to approximate the experience relatedness in the domain (Table 2): Top 5 concept pairs captured by BoW have an overall acceptance rate of more than 64% in different domains, which is at least 20% higher than SAMPO. The neural KGs have the best quality in all domains. However, there is no single knowledge that can lead to a neural KG that consistently has the highest quality across domains. K(SW) incorporating word association performed the best in the restaurant domain, while K(CM), incorporating physical common sense, had a 10-point more acceptance rate in the electronics domain than the others. The difference in the quality of these KGs is also confirmed by the results of MAP. Figure 2 compares the PR curves of these graphs, revealing that the basic and neural KGs constructed by CoNex have much higher precision and recall than SAMPO in all domains.

In general, K(BoW) does not consistently achieve better performance than BoW in different domains. This may be because transfer learning typically requires a sufficiently large and high-quality training set [Hwang et al., 2020]. As described above, we ensure the pairs in the training set are highly related by constraining the maximum ranking (top k)

4. Fleiss' Kappa scores for each annotated sets were 0.61 (hotel), 0.68 (restaurant), and 0.62 (electronics).

and the minimum similarity of concept pairs. The constraints in the hotel domain (top k < 10, similarity > 0.8) allow us to obtain a high-quality training set of 10K from the basic graph. However, in the restaurant domain, we need to lower the requirements (top $k < 30$, similarity > 0.8) to obtain a training set of the same size because this domain has fewer highly associated experiential factors. In addition, the same quality control (top $k < 30$, similarity > 0.8) in the electronics domain can only result in a 4K training set. Our results show that the size and quality of the latter two training sets can not benefit the language model. In this case, CoNex augments the training set by introducing suitable external data, enabling the language model to transfer knowledge successfully.

The fact that K(SAMPO) did not perform better than K(BOW) across domains indicates that domain-specific knowledge of opinion implications (i.e., data from SAMPO) did not improve neural KGs' knowledge prediction quality. This suggests that SAMPO's approach to this task does not provide more domain information than ours.

3.2 Task-based Evaluation

Identifying the experience-oriented relationship between concepts can benefit many downstream applications. For example, by knowing the range of factors that may influence a customer's experience with a product/service, the provider can respond more appropriately and recommend the customer's inquiry. Some of the prior related works have directly targeted such applications (e.g., [Xian et al., 2019, Kanouchi et al., 2020, Bjerva et al., 2020]. To showcase the potential of our research outcome to enhance real-world solutions in comparison with existing approaches, we design an answer retrieval task for the application of experience-oriented search.

We choose SubjQA [Bjerva et al., 2020] as the dataset for this experiment. SubjQA contains crowdsourced human labels of whether a user review in its pool contains the answers to various subjective queries around different factors concerning customer experiences. For example, the question "Was the service good?" inquires about users' experience with the hotel service. The sentence "The front desk was super helpful and friendly." in the review is marked as a valid answer, while the review "It's a short walk from the Yonge and Bloor station." is not (labeled as 'answer-not-found').

This dataset could serve as ground truth for evaluating IR-based solutions that respond to experience-oriented queries. Existing systems work well in factual search tasks but tend to struggle with such subjective queries [Bjerva et al., 2020].

Description of Answer Retrieval Task The task is to identify as many answers from a candidate pool as possible in response to a query of some experiential factor. We create a pool of positive and negative candidates for each query from SubjQA. Positive candidates are phrases or sentences marked as *answer*s in SubjQA, while negative candidates are reviews marked as *answer-not-found*. We selected three domains from the SubjQA dataset for this task: hotel (171 queries and 3,434 candidates), restaurant (238 queries and 3,379 candidates), and electronics (314 queries and 3,324 candidates).

Settings We create an out-of-the-box classifier using BERT Score [Zhang et al., 2019] to measure the similarities between the query and a candidate; if the score exceeds a certain threshold θ, the candidate will be labeled as positive (i.e., is an answer to the input query),

	Hotel			Restaurant			Electronics		
	$\theta = 0.85, 0.86, 0.87$			$\theta = 0.85, 0.86, 0.87$			$\theta = 0.85, 0.86, 0.87$		
No-feature	20.8	15.0	10.4	11.3	8.0	5.9	13.2	8.5	5.6
GloVe	24.0	16.4	11.1	15.5	10.6	7.6	16.0	10.4	6.5
SAMPO	30.5	22.7	15.6	17.0	12.0	8.9	16.2	11.4	7.4
BoW	37.7	30.1	22.0	22.0	14.3	10.5	**26.6**	16.7	12.1
K(BoW)	39.2	29.1	20.4	23.4	18.6	14.3	23.4	17.1	11.8
K(SAMPO)	37.0	29.1	19.3	24.9	18.3	12.7	23.6	19.1	12.5
K(SW)	40.1	**30.7**	21.9	**29.7**	**23.7**	**17.5**	24.4	18.5	13.1
K(CM)	**40.4**	30.3	**24.2**	26.2	20.1	14.0	**26.6**	**20.9**	**14.3**

Table 3: Average F1 scores of classifiers empowered by different knowledge in answer retrieval tasks across domains; θ represents the classification threshold.

otherwise negative (i.e., not an answer). In practice, the classifier generally produces values within a narrow range[5]. For example, in our cases, 0.85 and 0.87 are close to the min and max values across domains. That is, the score will be near 0.87 if two sentences are highly relevant, but even if they are not particularly relevant, the score is still close to 0.85. Therefore, we report the results with different thresholds ranging from 0.85 to 0.87.

Using Experience Knowledge Graph as Tool The experience KG can augment the classifier by enriching each input query with the top k ($k = 5$) related concepts to the experiential factor mentioned in the query. These concepts are also used as queries in the classifier, just as users enter a variety of keywords to get more relevant results when searching for an experience. If these keywords indeed encode new factors related to the queried experience, the classifier would be able to identify more true positives. However, if augmented keywords are actually unrelated, it would lead to an increase in false positives.

Baselines On this task, we evaluate the performance of six experience KGs as described in Section 3, i.e., **SAMPO**, **BoW**, **K(BoW)**, **K(SAMPO)**, **K(SW)**, and **K(CM)**. Additionally, to compare our proposed experiential relatedness features with the conventional semantic similarity features, we also use Gensim[6] to get the five most similar words in **GloVe** embedding to the experiential factor identified in a query to augment the original input, which serves as another baseline method.

Result Here we report the average F1 scores across all sampled queries to assess the performance of different classifiers built based on the features from the aforementioned methods. The average F1 scores of the classifiers empowered by different KGs in each domain are reported in Table 3. The "no-feature" classifier is the baseline using only the query to identify answers from candidates. The results show that the neural KG-based classifiers performed the best on all domains, achieving F1 scores at least 7 points higher than the "no-feature" model. The GloVe-based classifiers' performance was inferior to that

5. https://github.com/Tiiiger/bert_score/blob/master/journal/rescale_baseline.md
6. https://radimrehurek.com/gensim/

of other experience KG-based models, suggesting that experientially related concepts can better fuel experience-oriented search tasks than semantically related concepts. Results from the task-based evaluation were consistent with those of human evaluation in Section 3. For example, in the restaurant domain, the K(SW)-based model performed best, while in the electronics domain, the K(CM)-based model performed best. We suspect that these differences in performance are due to the SWOW training set being more subjective than the CM ones (i.e., common factual knowledge), resulting in SWOW being more useful in boosting models in domains with a higher level of subjectivity and CM assisting in the opposite way. The subjectivity of restaurant domains may be empirically measured by calculating the average number of experiential concepts (525K/144M=0.36), which is significantly higher than the electronics domain (822K/874M=0.09). Hotel domains have a medium level of subjectivity when compared to other twos (216K/155M=0.14), which may explain why K(SW) and K(CM) perform closely on it. This explanation may be further tested by comparing K(CM) and K(SWOW)'s performance across more domains.

4. Related Work

Modeling Consumer Experience Many previous studies on consumer experiences have analyzed user-generated content to extract and summarize experience-related opinions [Hu and Liu, 2004, Poria et al., 2016] and sentiments [Blair-Goldensohn et al., 2008, Diao et al., 2014]. However, the focus of this paper is to build on the identified experiences to further uncover the subjective association between experiential factors. To our knowledge, SAMPO [Bhutani et al., 2020] is the only study that looks specifically into subjective relationships to capture the implied connections between consumers' opinions (i.e., judgments about an object). An example of their results is that "good, drink" implies "great, bar". Such experientially related opinions show promise in many real-world applications, such as providing explainable suggestions for vague requests [Kanouchi et al., 2020] and answering users' subjective questions [Bjerva et al., 2020]. Unlike SAMPO, our work focuses only on the factors that people experience without being limited by their views of those factors, and our experimental results suggest that SAMPO's approach cannot be effectively applied to capture experientially related factors.

Knowledge Extraction Previous research has investigated acquiring knowledge with factual relationships using expert knowledge [Lenat, 1995, Miller, 1995], semi-structured text extraction [Auer et al., 2007, Suchanek et al., 2007], and unstructured text extraction [Dong et al., 2014] in both general and commercial domains [Dong, 2018, Xu et al., 2020]. The outcome knowledge graphs can be used to support downstream applications such as domain-specific recommendations [Xian et al., 2019]s and question-answering [Yang et al., 2017]. In recent years, pre-trained language models also show promise in generating hypothetical factual knowledge [Bosselut et al., 2019, Petroni et al., 2019, Hwang et al., 2020], complementing the knowledge coverage of traditional methods. However, the experiential relationship we focus on is generally subjective and implicit and cannot be sufficiently explained by objective factual relationships. Existing works on the extraction of implicit associations between concepts focus primarily on computing semantic similarity and relatedness [Gabrilovich et al., 2007, Agirre et al., 2009, Radinsky et al., 2011] between words. While semantic relations indicate how words are associated with meanings in general

domains [Zhang et al., 2013], they do not necessarily reflect the interconnection between consumer experience-related factors within a particular domain. For example. *cleanliness* and *room smell* are associated experientially but not semantically.

5. Conclusion

This paper introduces a novel relationship called experiential relatedness to represent the domain-specific association between factors concerning consumer experiences. We propose an automated and generalizable pipeline, CoNex, for building knowledge graphs that map out consumer experiences in specific domains. Our study shows that implicit experiential relations can be captured from user-generated content by examining the distribution of experience-related concepts. In addition, language models fine-tuned on domain and external knowledge can learn experiential connections between concepts and infer novel and accurate experience-related knowledge. Future work can explore how these constructed experience KGs can be better applied to many real-world experience-related applications, such as experiential search [Evensen et al., 2019] and explainable recommendations [Kanouchi et al., 2020].

Acknowledgements

We are very grateful to Alon Halevy for his invaluable advice and assistance throughout this project. This work is partially supported by the Research Grants Council of the Hong Kong Special Administrative Region under General Research Fund (GRF) with Grant No. 16203421.

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Pasca, and Aitor Soroa. A study on similarity and relatedness using distributional and wordnet-based approaches. 2009.

Joseph Alba, John Lynch, Barton Weitz, Chris Janiszewski, Richard Lutz, Alan Sawyer, and Stacy Wood. Interactive home shopping: consumer, retailer, and manufacturer incentives to participate in electronic marketplaces. *Journal of marketing*, 61(3):38–53, 1997.

Eric J Arnould. *Consumers*. McGraw-Hill/Irwin series in marketing. McGraw-Hill/Irwin, Boston, 2nd ed.. edition, 2004. ISBN 0072537140.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer, 2007.

Nikita Bhutani, Aaron Traylor, Chen Chen, Xiaolan Wang, Behzad Golshan, and Wang-Chiew Tan. Sampo: Unsupervised knowledge base construction for opinions and implications. In *Automated Knowledge Base Construction*, 2020.

Johannes Bjerva, Nikita Bhutani, Behzad Golshan, Wang-Chiew Tan, and Isabelle Augenstein. Subjqa: A dataset for subjectivity and review comprehension. *arXiv preprint arXiv:2004.14283*, 2020.

Sasha Blair-Goldensohn, Kerry Hannan, Ryan McDonald, Tyler Neylon, George Reis, and Jeff Reynar. Building a sentiment summarizer for local service reviews. 2008.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. Comet: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, 2019.

J Joško Brakus, Bernd H Schmitt, and Shi Zhang. Experiential product attributes and preferences for new products: The role of processing fluency. *Journal of Business Research*, 67(11):2291–2298, 2014.

Simon De Deyne, Amy Perfors, and Daniel J Navarro. Predicting human similarity judgments with distributional models: The value of word associations. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers*, pages 1861–1870, 2016.

Simon De Deyne, Danielle J Navarro, Amy Perfors, Marc Brysbaert, and Gert Storms. The "small world of words" english word association norms for over 12,000 cue words. *Behavior research methods*, 51(3):987–1006, 2019.

Qiming Diao, Minghui Qiu, Chao-Yuan Wu, Alexander J Smola, Jing Jiang, and Chong Wang. Jointly modeling aspects, ratings and sentiments for movie recommendation (jmars). In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 193–202, 2014.

Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 601–610, 2014.

Xin Luna Dong. Challenges and innovations in building a product knowledge graph. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2869–2869, 2018.

Sara Evensen, Aaron Feng, Alon Halevy, Jinfeng Li, Vivian Li, Yuliang Li, Huining Liu, George Mihaila, John Morales, Natalie Nuno, et al. Voyageur: An experiential travel search engine. In *The World Wide Web Conference*, pages 3511–5, 2019.

Evgeniy Gabrilovich, Shaul Markovitch, et al. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJcAI*, volume 7, pages 1606–1611, 2007.

Alon Y Halevy. The ubiquity of subjectivity. *IEEE Data Eng. Bull.*, 42(1):6–9, 2019.

Catherine Havasi, Robert Speer, and Jason Alonso. Conceptnet 3: a flexible, multilingual semantic network for common sense knowledge. In *Recent advances in natural language processing*, pages 27–29. John Benjamins Philadelphia, PA, 2007.

Ruining He and Julian McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, pages 507–517, 2016.

Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177, 2004.

Peng Huang, Nicholas H Lurie, and Sabyasachi Mitra. Searching for experience on the web: An empirical examination of consumer behavior for search and experience goods. *Journal of marketing*, 73(2):55–69, 2009.

Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. *arXiv preprint arXiv:2010.05953*, 2020.

Shin Kanouchi, Masato Neishi, Yuta Hayashibe, Hiroki Ouchi, and Naoaki Okazaki. You may like this hotel because...: Identifying evidence for explainable recommendations. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 890–899, 2020.

Lisa R Klein. Evaluating the potential of interactive media through a new lens: Search versus experience goods. *Journal of business research*, 41(3):195–203, 1998.

Douglas B Lenat. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38, 1995.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, 2020.

Yuliang Li, Aaron Feng, Jinfeng Li, Saran Mumick, Alon Halevy, Vivian Li, and Wang-Chiew Tan. Subjective databases. *Proceedings of the VLDB Endowment*, 12(11):1330–1343, 2019.

Diego Marcheggiani, Oscar Täckström, Andrea Esuli, and Fabrizio Sebastiani. Hierarchical multi-label conditional random fields for aspect-oriented opinion mining. In *European conference on information retrieval*, pages 273–285. Springer, 2014.

George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38 (11):39–41, 1995.

Phillip Nelson. Information and consumer behavior. *Journal of political economy*, 78(2): 311–329, 1970.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? 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 2463–2473, 2019.

Soujanya Poria, Erik Cambria, and Alexander Gelbukh. Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems*, 108:42–49, 2016.

Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th international conference on World wide web*, pages 337–346, 2011.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67, 2020.

Eldar Shafir, Itamar Simonson, and Amos Tversky. Reason-based choice. *Cognition*, 49 (1-2):11–36, 1993.

Itamar Simonson. Choice based on reasons: The case of attraction and compromise effects. *Journal of consumer research*, 16(2):158–174, 1989.

Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706, 2007.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Q Zhu. Probase: A probabilistic taxonomy for text understanding. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pages 481–492, 2012.

Yikun Xian, Zuohui Fu, Shan Muthukrishnan, Gerard De Melo, and Yongfeng Zhang. Reinforcement knowledge graph reasoning for explainable recommendation. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 285–294, 2019.

Da Xu, Chuanwei Ruan, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. Product knowledge graph embedding for e-commerce. In *Proceedings of the 13th international conference on web search and data mining*, pages 672–680, 2020.

Shuo Yang, Lei Zou, Zhongyuan Wang, Jun Yan, and Ji-Rong Wen. Efficiently answering technical questions—a knowledge graph approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

Ziqi Zhang, Anna Lisa Gentile, and Fabio Ciravegna. Recent advances in methods of lexical semantic relatedness-a survey. *Natural Language Engineering*, 19(4):411, 2013.

Appendix A. Appendix

A.1 Implementation Details about BART

We use the BART-base version of the model[7], which has 12 layers, 768-dimensional hidden states, 16 attention heads in its self-attention layers, and 139M total parameters. The batch size is 32, and the learning rate is 2e-5. Each model is trained for 30 epochs, and then the best model is saved by monitoring the minimum validation loss. To increase the robustness of the results, for each neural KG in the experiment, we use three random seeds for its training set partitioning and model training. We collected a total of 15 concepts from the top 5 outputs of the three models and then sampled 5 of them according to the frequency of the concepts in them as the final top 5 outputs.

A.2 Examples

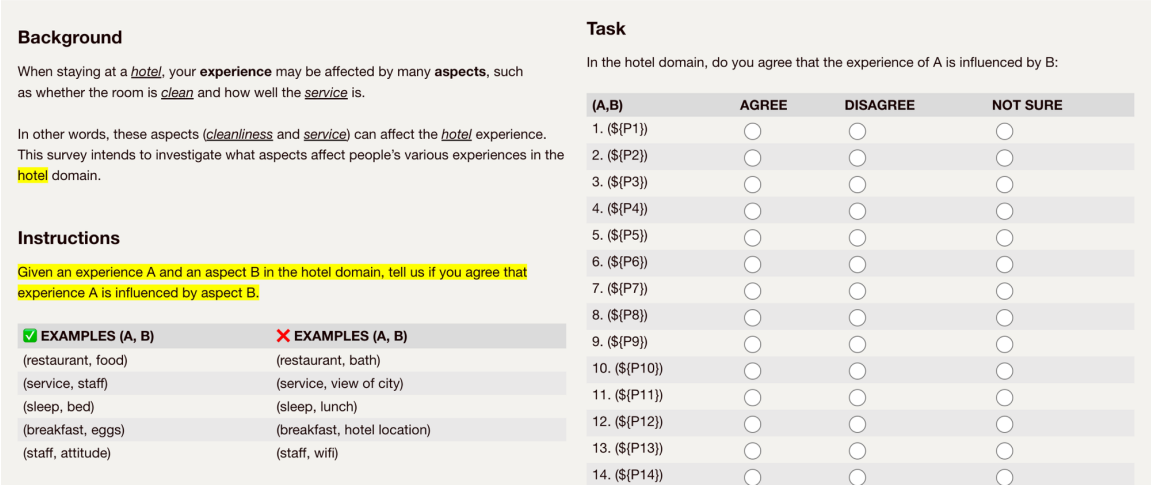
Domain	Acceptance	Concept A	Concept B
Hotel	✓	`sleep quality`	`mattress pad`
	✓	`fitness center`	`exercise facility`
	✓	`staff member`	`english speak`
	✗	`meal`	`sleep`
	✗	`sofa`	`work desk`
Restaurant	✓	`lunch`	`food`
	✓	`restaurant`	`cuisine`
	✓	`service`	`delivery`
	✗	`waiter`	`atmosphere`
	✗	`onion ring`	`shoe string`
Electronics	✓	`music`	`quality sound`
	✓	`seller`	`service`
	✓	`color`	`shade`
	✗	`card`	`angle`
	✗	`customer service`	`print`

Table 4: Examples of acceptable (✓) and unacceptable (✗) experiential related concepts mined by our KGs.

A.3 MTurk HIT Design

We designed a human intelligence task (HIT) on MTurk to evaluate the quality of the experience-related concepts mined by our KGs. A snippet of a HIT is shown in Figure 3. For each HIT, workers received instruction on five each positive and negative examples. They were asked to annotate 20 pairs of concepts ("*Do you agree that experience A is influenced by aspect B*"). Each question consists of three options (agree, disagree, not sure). Five pairs of concepts pre-annotated by researchers were included in the questions and

7. from HuggingFace's implementation [Wolf et al., 2019]



Background

When staying at a *hotel*, your **experience** may be affected by many **aspects**, such as whether the room is *clean* and how well the *service* is.

In other words, these aspects (*cleanliness* and *service*) can affect the *hotel* experience. This survey intends to investigate what aspects affect people's various experiences in the hotel domain.

Instructions

Given an experience A and an aspect B in the hotel domain, tell us if you agree that experience A is influenced by aspect B.

✅ EXAMPLES (A, B)	❌ EXAMPLES (A, B)
(restaurant, food)	(restaurant, bath)
(service, staff)	(service, view of city)
(sleep, bed)	(sleep, lunch)
(breakfast, eggs)	(breakfast, hotel location)
(staff, attitude)	(staff, wifi)

Task

In the hotel domain, do you agree that the experience of A is influenced by B:

(A,B)	AGREE	DISAGREE	NOT SURE
1. (${P1})	○	○	○
2. (${P2})	○	○	○
3. (${P3})	○	○	○
4. (${P4})	○	○	○
5. (${P5})	○	○	○
6. (${P6})	○	○	○
7. (${P7})	○	○	○
8. (${P8})	○	○	○
9. (${P9})	○	○	○
10. (${P10})	○	○	○
11. (${P11})	○	○	○
12. (${P12})	○	○	○
13. (${P13})	○	○	○
14. (${P14})	○	○	○

Figure 3: The screenshot of the HIT design.

served as standard tests for assessing workers. We only kept the HIT results that passed the tests. A HIT generally took 2 to 8 minutes and was paid $1 (about $7.5/hour).