Learning What to Ask: Mining Product Attributes for E-commerce Sales from Massive Dialogue Corpora

Yan Fan
fanyan.fy@alibaba-inc.com
Alibaba Group
Hangzhou, China

Chengyu Wang
chengyu.wcy@alibaba-inc.com
Alibaba Group
Hangzhou, China

Fan Feng
fengfan.fengfan@alibaba-inc.com
Alibaba Group
Hangzhou, China

Hengbin Cui
alexcui.chb@alibaba-inc.com
Alibaba Group
Hangzhou, China

Yuchuan Wu
shengxiu.wyc@alibaba-inc.com
Alibaba Group
Hangzhou, China

Yongbin Li*
shuide.lyb@alibaba-inc.com
Alibaba Group
Hangzhou, China

ABSTRACT

Conversational Recommender Systems (CRSs) are extensively applied in e-commercial platforms that recommend items to users. To ensure accurate recommendation, agents usually ask for users' preferences towards specific product attributes which are pre-defined by humans. In e-commercial platforms, however, the number of products easily reaches to billions, making it prohibitive to predefine decisive attributes for efficient recommendation due to the lack of substantial human resources and the scarce domain expertise. In this work, we present MOSAIC, a novel knowledge mining and conversational assistance framework that extracts core product attributes from massive dialogue corpora for better conversational recommendation experience. It first extracts user-agent interaction utterances from massive corpora that contain product attributes. A Joint Attribute and Value Extraction (JAVE) network is designed to extract product attributes from user-agent interaction utterances. Finally, MOSAIC generates attribute sets that frequently appear in dialogues as the target attributes for agents to request, and serve as an assistant to guide the dialogue flow. To prove the effectiveness of MOSAIC, we show that it consistently improves the overall recommendation performance of our CRS system. An industrial demonstration scenario is further presented to show how it benefits online shopping experiences.

CCS CONCEPTS

• **Computing methodologies → Discourse, dialogue and pragmatics**; • **Information systems → Electronic commerce**.

KEYWORDS

conversational recommender system, product attribute mining, dialogue flow assistance, e-commerce application

CIKM '23, October 21–25, 2023, Birmingham, United Kingdom
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0124-5/23/10…$15.00
https://doi.org/10.1145/3583780.3614745

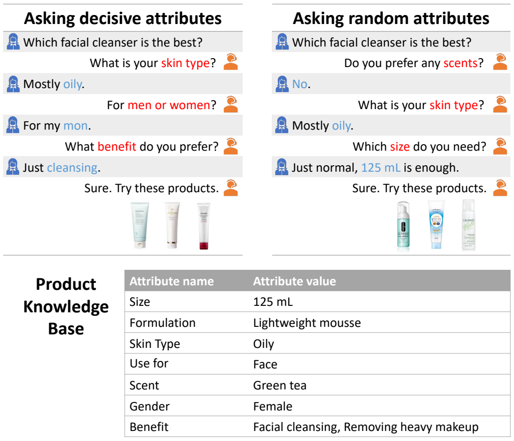
ACM Reference Format:
Yan Fan, Chengyu Wang, Fan Feng, Hengbin Cui, Yuchuan Wu, and Yongbin Li*. 2023. Learning What to Ask: Mining Product Attributes for E-commerce Sales from Massive Dialogue Corpora. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), October 21–25, 2023, Birmingham, United Kingdom.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3583780.3614745

1 INTRODUCTION

Conversational Recommender Systems (CRSs) are frequently applied in e-commercial platforms that recommend items to users via multiple turns of conversational interactions [6, 13, 14]. In CRSs, it is essential to understand the users' preferences in order to recommend the most suitable items, so that to increase product sales and improve user experience in online shopping [12].

In the literature, how to recommend items through multiple turns of conversations has often been identified as a core task in CRSs. The CRS system seeks to determine the next question to ask and acquires user preferences on certain attributes of products. It follows the "slot-filling" approach of collecting user information as in task-oriented dialogue systems [16], but is more flexible to select decisive attributes that reflect user preferences instead of filling a fixed collection of slots. A number of recent works model the task with reinforcement learning, to increase the recommendation efficiency with a minimum number of interaction turns [3, 7, 8]. A few works also identify attributes that help to narrow down the search space for item recommendation [2, 18]. However, in e-commercial platforms, we are facing a more challenging and open-world problem where the number of product categories easily is over ten thousand. Hence, it is extremely expensive and even prohibitive for domain experts to pre-define the decisive attribute sets and possible values for each product category. Consider the sales agent service AliMe[1] in Alibaba Group. AliMe produces over 10 millions of human-human dialogues each day, involving nearly twenty thousand product categories from dozens of domains. Over 60 percent of dialogues focus on pre-sale recommendation, where nearly 20 percent of them express explicit needs for certain product category. Asking the right questions helps to guide customers with strong willingness to purchase, leading to a significant increase of sales. A motivation example is also presented in Figure 1. Therefore, with massive dialogue corpora collected from AliMe available, a

[1] https://www.alixiaomi.com/



Asking decisive attributes

Which facial cleanser is the best?

What is your skin type?

Mostly oily.

For men or women?

For my mon.

What benefit do you prefer?

Just cleansing.

Sure. Try these products.

Asking random attributes

Which facial cleanser is the best?

Do you prefer any scents?

No.

What is your skin type?

Mostly oily.

Which size do you need?

Just normal, 125 mL is enough.

Sure. Try these products.

Product Knowledge Base

Attribute name	Attribute value
Size	125 mL
Formulation	Lightweight mousse
Skin Type	Oily
Use for	Face
Scent	Green tea
Gender	Female
Benefit	Facial cleansing, Removing heavy makeup

Figure 1: An example that compares two strategies of asking attributes for facial cleansers with product KBs containing attribute-value pairs.

natural question is: *how can we design a framework that can extract decisive attributes from massive dialogue corpora in order to guide the e-commercial CRS?*

In this paper, we present MOSAIC (**M**ining pr**O**duct attribute**S** m**A**ssive d**I**alogue **C**orpora), a novel knowledge mining and conversational assistance framework that extracts decisive attributes for any product categories and benefits the overall recommendation performance of the AliMe CRS. The high-level framework is illustrated in Figure 2. Specifically, our MOSAIC framework seamlessly integrates the following four modules:

- **Attribute-oriented Interaction Detection**: This module aims at extracting user-agent interaction utterances from massive dialogue corpora that provide strong signals of discussing certain product attributes.
- **Product Attribute Extraction**: In this module, we design the *Joint Attribute and Value Extraction* (JAVE) network to extract product attributes from user-agent interaction utterances. It employs multi-task training to detect both attribute names and their values from user-agent interactions.
- **Frequent Attribute Mining**: This module mines attribute sets that frequently appear in dialogues as the target attributes for the agents to request. It works efficiently using Apriori-based heuristics on distributed computing clusters.
- **Dialogue Flow Assistance**: MOSAIC serves as a useful tool to provide dialogue flow assistance for sellers. During each around of user-agent interaction, the agent asks the customer the preference of an attribute, and finally delivers item recommendations.

To verify the effectiveness of MOSAIC, we briefly report our in-house evaluation results and the overall recommendation performance of our CRS system. An industrial demonstration case is further presented, demonstrating how MOSAIC benefits online shopping experiences of both sellers and customers.

2 MOSAIC: PROPOSED FRAMEWORK

In this section, we present the technique details of the individual modules in the MOSAIC framework.

2.1 Attribute-oriented Interaction Detection

In the literature, a majority of research works focus on the classification of user intents, such as [1, 5, 15, 17]. In our work, however, we are more interested in how the user and the agent interact with each other w.r.t. specific product attributes. Formally, given a dialogue session $S = \{A_1, U_1, \cdots, A_i, U_i, \cdots\}$[2] where A_i and U_i represent the i-th agent/user utterances, respectively. We regard the pair $< A_i, U_i >$ as a *Candidate Attribute-oriented Interaction* (CAI) if the act of A_i is *Asking for Preferences* and the the act of U_i is *Providing Preferences*. In this step, although we do not know the exact product attributes, the collections of CAI pairs provide strong signals that the agent and the user are discussing some product attributes. In the implementation, we train two BERT-based classifiers [4] to classify agent and user acts, respectively. The collection of all agent acts includes *Request Attributes*, *Recommend Products*, and *Others*, while the collection of user acts includes *Provide Categories*, *Provide Preferences*, and *Others*. Note that we do not employ any fine-grained taxonomies of dialogue acts here to simplify the data labeling and training process, as well as to increase the robustness of the obtained classifiers.

2.2 Product Attribute Extraction

After the detection of CAI pairs, we proceed to detect attribute names. Consider the following CAI pair from the cosmetics domain:

Agent: What is your skin type? **User:** Mostly oily.

Here, the attribute name ("skin type") co-occurs with the value ("oily") within the same CAI pair. A native approach to this task is to leverage the zero-shot power of Large Language Models (LLMs) such as ChatGPT[3]. However, we observe that it can have poor abilities for extracting structured data (see CASE I presented in Figure 3 where the attribute name "neutral" should be extracted). Even when all candidate answers are provided in the prompts (for example, "oily", "neutral" and "dry"), ChatGPT sometimes fails to understand the dialogue semantics (see CASE II where "not very oily" should be mapped to "neutral").

In our work, we extend [16] and design the *Joint Attribute and Value Extraction* (JAVE) network for attribute extraction. The model architecture is illustrated in Figure 4. In order to reduce manual labeling in the cold-start stage, we utilize external product knowledge bases to automatically tag attribute names and corresponding values mentioned in CAI pairs. Here, we assume that when an attribute name and its value co-occur in a CAI pair, it is regarded as a positive sample for the JAVE network, based on the similar assumption in distant supervision [10].

In JAVE, we employ BERT [4] as the utterance-pair encoder to learn the representations of CAI pairs. A slot value generator decodes independently for all the possible domain-attribute pairs. This lite text decoder is built upon the encoder to generate possible corresponding attribute value for every pair. Meanwhile, the slot gate predicts whether the current domain-slot pair is triggered by the CAI pair and the logical relationship (e.g., equals to or not equals to) implied in the context. Decode the sequence-to-sequence

[2] Here, we only focus on how the agent interacts with the user before item recommendation. Other types of dialogues such as chitchats and after-sales consulting are filtered by heuristics.

[3] https://chat.openai.com/

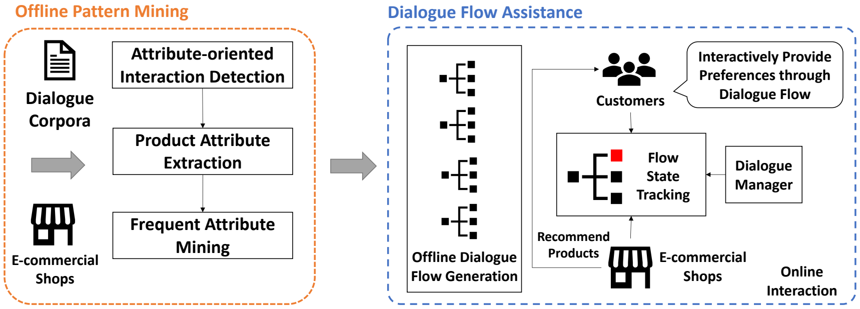
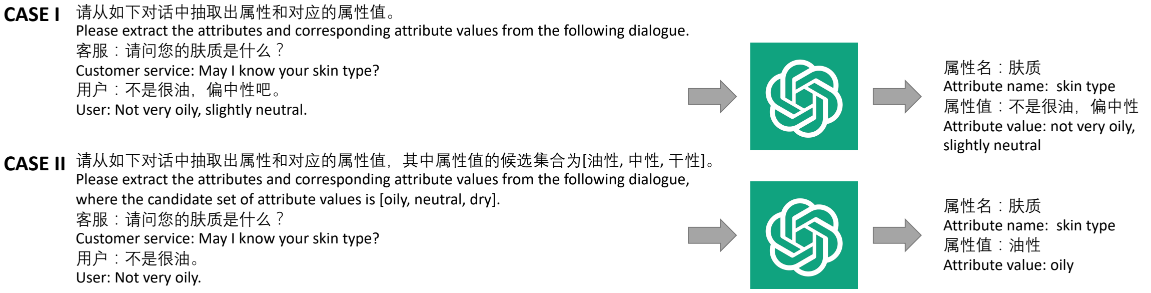

Offline Pattern Mining
Dialogue Corpora
E-commercial Shops
Attribute-oriented Interaction Detection
Product Attribute Extraction
Frequent Attribute Mining
Dialogue Flow Assistance
Offline Dialogue Flow Generation
Customers
Interactively Provide Preferences through Dialogue Flow
Flow State Tracking
Dialogue Manager
Recommend Products
E-commercial Shops
Online Interaction

Figure 2: The high-level framework of MOSAIC. Our first three steps (Attribute-oriented Interaction Detection, Product Attribute Extraction and Frequent Attribute Mining) mine product attributes for better recommendation. The Dialogue Flow Assistance module produces dialogue flow templates offline, and guides the interaction process online.



CASE I 请从如下对话中抽取出属性和对应的属性值。
Please extract the attributes and corresponding attribute values from the following dialogue.
客服：请问您的肤质是什么？
Customer service: May I know your skin type?
用户：不是很油，偏中性吧。
User: Not very oily, slightly neutral.

属性名：肤质
Attribute name: skin type
属性值：不是很油，偏中性
Attribute value: not very oily, slightly neutral

CASE II 请从如下对话中抽取出属性和对应的属性值，其中属性值的候选集合为[油性, 中性, 干性]。
Please extract the attributes and corresponding attribute values from the following dialogue, where the candidate set of attribute values is [oily, neutral, dry].
客服：请问您的肤质是什么？
Customer service: May I know your skin type?
用户：不是很油。
User: Not very oily.

属性名：肤质
Attribute name: skin type
属性值：油性
Attribute value: oily

Figure 3: Cases of results generated by ChatGPT (GPT3.5-turbo) for product attribute extraction. The inputs and outputs are in Chinese and have been translated into English for reference.

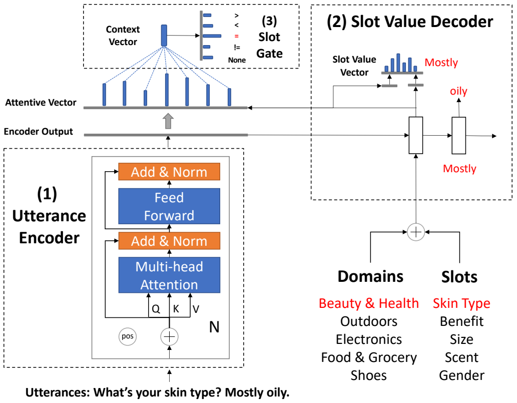


Figure 4: The JAVE model architecture.

(seq2seq) loss of the text decoder as $\mathcal{L}_{Decoder}$, and the cross-entropy loss of the slot gate as $\mathcal{L}_{Gate}$. The overall loss function of the JAVE model (denoted as $\mathcal{L}$) is formulated as:

$$\mathcal{L} = \mathcal{L}_{Gate} + \lambda \cdot \mathcal{L}_{Decoder} \tag{1}$$

where λ is the balancing hyper-parameter ($0 < \lambda < 1$). Based on our evaluation, the multi-task training technique improves the performance of attribute extraction by a large margin.

2.3 Frequent Attribute Mining

From previous modules, we can extract the product attributes. However, it is infeasible to ask questions about all the attributes that belong to the certain product category. This is because the collection can be very large due to the "long-tail" data distribution of the extracted results and the existence of inevitable noise.

This module further mines attribute sets that frequently appear in dialogues. Given a specific product category, denote C as a collection of the combination of arbitrary attribute name c, and $k(C)$ as the frequency that attributes co-occur among all the dialogue sessions. We regard C^* to be the best attribute set if the score of C^* (denoted as $s(C^*)$) is the highest among all the attribute sets:

$$s(C^*) = k(C^*) \cdot \log_M |C^*| \tag{2}$$

where $M > 1$ is a pre-defined constant. The scoring function $s(C^*)$ considers both the frequency and the size of attribute collections, in order to select the most suitable set for the agent to request. In our system, as the numbers of dialogue sessions can easily reach to billions, we mine frequent attribute sets based on Apriori-based heuristics [9] on distributed computing clusters.

2.4 Dialogue Flow Assistance

After we extract the frequent attribute sets, we need to help sellers construct the dialogue flow for effective recommendation. Here, we take "Essence" for example. MOSAIC generates frequent attribute sets {skin_type, benefit, gender} and possible value sets {oily, dry, neutral}, {anti-aging, moisturizing}, {male, female}. Each product category is associated with frequent attribute-value pairs and several dialogue sessions containing CAI pairs as expert knowledge. Sellers are provided with above information so that they can customize their own dialogue flow via a tree-based XMind plugin, as shown

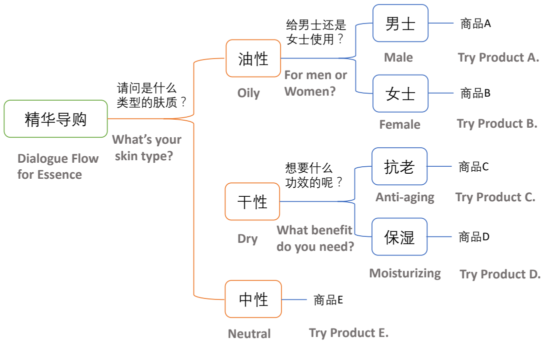


Figure 5: Customizing dialogue flows for sellers. The sample product category here is “Essence”.

Model	Macro-F1	Micro-F1	Weighted-F1
Agent Model	0.8935	0.8911	0.8911
User Model	0.9275	0.9376	0.9377

Table 1: The performance of interaction detection classifiers in terms of F1.

Method	Accuracy	Improvement
BERT	0.56	-
JAVE (Ours)	**0.78**	**+22%**

Table 2: The online deployment performance (A/B test) of JAVE in terms of accuracy.

in Figure 5. Here, each node represents an attribute question to for the seller to acquire, an option of all possible attribute values, or a product recommendation result.

In the online recommendation stage, we utilize the generated flow to manage the dialogue. During each around of user-agent interaction, the agent asks the customer the preference of an attribute from the selected attribute collection. After a certain around of interactions, our system retrieves the most suitable items for recommendation. Hence, MOSAIC serves as a useful tool to provide dialogue flow assistance for sellers.

3 BRIEF EVALUATION RESULTS

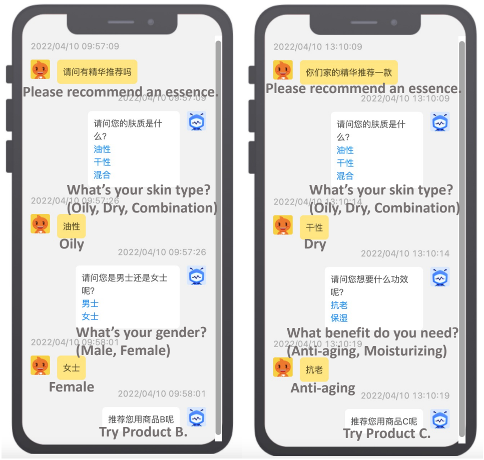
In this section, we briefly report our in-house evaluation results of the proposed MOSAIC framework.

3.1 Effectiveness of Individual Models

As we focus on the effectiveness of our framework for real-world applications, we specifically show how our models perform over our AliMe dialogue data. For interaction detection, to meet the high QPS (Query Per Second) demands online, we leverage the tiny version of the BERT model from [11] as our backbone, which has two attention heads, two hidden layers and the hidden size of 128. The datasets of agent and user utterances are around 10k and 3k, respectively, which are further split into training, validation and testing sets. The results of both classifiers are shown in Table 1. We can see that we can achieve satisfactory performance using very small models. As for the attribute extraction model, we conduct online A/B tests, where the baseline is the vanilla BERT model.

Category	w/o MOSAIC	w/ MOSAIC	Improv.
Clothes	0.1987	0.2339	3.52%
Shoes	0.1868	0.2010	1.42%
Outdoors	0.1480	0.2445	10.65%
Food & Grocery	0.2481	0.3395	9.14%
Electronics	0.1359	0.2111	7.52%
General Merchandise	0.1961	0.2698	7.37%
Automotive Parts	0.4054	0.4400	3.46%

Table 3: The online performance (A/B test) of sellers from seven industry categories in terms of ICR.


2022/04/10 09:57:09
请问有精华推荐吗
Please recommend an essence.
2022/04/10 09:57:09
请问您的肤质是什么?
油性
干性
混合
What’s your skin type?
(Oily, Dry, Combination)
2022/04/10 09:57:26
油性
Oily
2022/04/10 09:57:26
请问您是男士还是女士呢?
男士
女士
What’s your gender?
(Male, Female)
2022/04/10 09:58:01
女士
Female
2022/04/10 09:58:01
推荐您用商品B呢
Try Product B.

2022/04/10 13:10:09
你们家的精华推荐一款
Please recommend an essence.
2022/04/10 13:10:09
请问您的肤质是什么?
油性
干性
混合
What’s your skin type?
(Oily, Dry, Combination)
2022/04/10 13:10:14
干性
Dry
2022/04/10 13:10:14
请问您想要什么功效呢?
抗老
保湿
What benefit do you need?
(Anti-aging, Moisturizing)
2022/04/10 13:10:19
抗老
Anti-aging
2022/04/10 13:10:19
推荐您用商品C呢
Try Product C.

Figure 6: Demonstration screenshots of online dialogues. Due to trademark protection, we have removed specific brands and product names in the figure.

From the results in Table 2, it is clearly proven that our model improves the extraction performance by a large margin (22% in terms of accuracy).

3.2 Overall Recommendation Evaluation

After internal evaluation, we have deployed our framework online, providing service for over hundreds of sellers and their customers. To show that the MOSAIC framework can improve the overall item recommendation performance, we report the Inquiry Conversion Rate (ICR) of our CRS with and without the MOSAIC framework, which is the number of conversions (i.e., purchasing the products that we recommend) divided by the total number of visitors. In Table 3, we list the ICR results of sellers from seven industry categories (e.g., Food & Grocery, Outdoors, Clothes). The results show that improvement of ICR is highly consistent across sellers in different categories, leading to better e-commerce sales.

4 DEMONSTRATION SCREENSHOTS

The functionalities of MOSAIC have been fully integrated into the AliMe customer service for Taobao.com, one of the largest e-commerce platform in China. Some screenshots showing how MOSAIC works are shown in Figure 6. As seen, it provides clear and explicit guidance for both sellers and customers to find the best product satisfying customers’ needs.

REFERENCES

[1] Wanling Cai and Li Chen. 2019. Towards a Taxonomy of User Feedback Intents for Conversational Recommendations. In *RecSys (Late-Breaking Results)*, Vol. 2431. 51–55.

[2] Sapna Ria Chakraborty, Anagha M., Kartikeya Vats, Khyati Baradia, Tanveer Khan, Sandipan Sarkar, and Sujoy Roychowdhury. 2019. Recommendence and Fashionsence: Online Fashion Advisor for Offline Experience. In *COMAD/CODS*. 256–259.

[3] Zhendong Chu, Hongning Wang, Yun Xiao, Bo Long, and Lingfei Wu. 2023. Meta Policy Learning for Cold-Start Conversational Recommendation. In *WSDM*. 222–230.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*. 4171–4186.

[5] Yan Fan, Chengyu Wang, Peng He, and Yunhua Hu. 2022. Building Multi-turn Query Interpreters for E-commercial Chatbots with Sparse-to-dense Attentive Modeling. In *WSDM*. 1577–1580.

[6] Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2021. A Survey on Conversational Recommender Systems. *ACM Comput. Surv.* 54, 5 (2021), 105:1–105:36.

[7] Yucheng Jin, Wanling Cai, Li Chen, Nyi Nyi Htun, and Katrien Verbert. 2019. MusicBot: Evaluating Critiquing-Based Music Recommenders with Conversational Interaction. In *CIKM*. 951–960.

[8] Heeseon Kim, Hyeongjun Yang, and Kyong-Ho Lee. 2023. Confident Action Decision via Hierarchical Policy Learning for Conversational Recommendation. In *WWW*. 1386–1395.

[9] Jure Leskovec, Anand Rajaraman, and Jeffrey D. Ullman. 2014. *Mining of Massive Datasets, 2nd Ed.* Cambridge University Press.

[10] Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *ACL*, Keh-Yih Su, Jian Su, and Janyce Wiebe (Eds.). 1003–1011.

[11] Minghui Qiu, Peng Li, Chengyu Wang, Haojie Pan, Ang Wang, Cen Chen, Xianyan Jia, Yaliang Li, Jun Huang, Deng Cai, and Wei Lin. 2021. EasyTransfer: A Simple and Scalable Deep Transfer Learning Platform for NLP Applications. In *CIKM*. 4075–4084.

[12] Filip Radlinski, Craig Boutilier, Deepak Ramachandran, and Ivan Vendrov. 2022. Subjective Attributes in Conversational Recommendation Systems: Challenges and Opportunities. In *AAAI*. 12287–12293.

[13] Yueming Sun and Yi Zhang. 2018. Conversational Recommender System. In *SIGIR*. 235–244.

[14] Dai Hoang Tran, Quan Z. Sheng, Wei Emma Zhang, Salma Abdalla Hamad, Munazza Zaib, Nguyen H. Tran, Lina Yao, and Nguyen Lu Dang Khoa. 2020. Deep Conversational Recommender Systems: A New Frontier for Goal-Oriented Dialogue Systems. *CoRR* abs/2004.13245 (2020).

[15] Chengyu Wang, Haojie Pan, Yuan Liu, Kehan Chen, Minghui Qiu, Wei Zhou, Jun Huang, Haiqing Chen, Wei Lin, and Deng Cai. 2021. MeLL: Large-scale Extensible User Intent Classification for Dialogue Systems with Meta Lifelong Learning. In *KDD*. ACM, 3649–3659.

[16] Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. 2019. Transferable Multi-Domain State Generator for Task-Oriented Dialogue Systems. In *ACL*. 808–819.

[17] Zhao Yan, Nan Duan, Peng Chen, Ming Zhou, Jianshe Zhou, and Zhoujun Li. 2017. Building Task-Oriented Dialogue Systems for Online Shopping. In *AAAI*. 4618–4626.

[18] Zi Yin, Keng-hao Chang, and Ruofei Zhang. 2017. DeepProbe: Information Directed Sequence Understanding and Chatbot Design via Recurrent Neural Networks. In *KDD*. 2131–2139.