An Automatic Procedure for Generating Datasets for Conversational Recommender Systems

Alessandro Suglia[1], Claudio Greco[1], Pierpaolo Basile[1], Giovanni Semeraro[1], and Annalina Caputo[2]

[1] Department of Computer Science, University of Bari Aldo Moro
[2] ADAPT Centre, Trinity College Dublin, Dublin, Ireland
`alessandro.suglia@gmail.com, claudiogaetanogreco@gmail.com,`
`{firstname.lastname}@uniba.it,annalina.caputo@adaptcentre.ie`

Abstract. Conversational Recommender Systems assist online users in their information-seeking and decision making tasks by supporting an interactive process with the aim of finding the most appealing items according to the user preferences. Unfortunately, collecting dialogues data to train these systems can be labour-intensive, especially for data-hungry Deep Learning models. Therefore, we propose an automatic procedure able to generate plausible dialogues from recommender systems datasets.

People have information needs of varying complexity, which can be solved by an intelligent agent able to answer questions formulated in a proper way, eventually considering user context and preferences. *Conversational Recommender Systems (CRS)* assist online users in their information-seeking and decision making tasks by supporting an interactive process [1] with the aim of finding the most appealing items according to the user preferences.

Unfortunately, collecting dialogues data required for the training phase of these systems can be really labour-intensive, especially for the latest data-hungry *Deep Learning* models. For this reason, synthetic dialogue datasets can be extremely useful in order to bootstrap effective dialogue systems able to support a goal-oriented conversation with the user. Therefore, we propose an automatic procedure able to generate plausible dialogues directly from well-known recommender systems datasets exploiting data coming from the Linked Open Data Cloud and contextual information related to the user.

Given a user u and his/her set of *binary* preferences, we trained a decision tree from the user u preferences expressed towards items represented using *Linked Open Data* binary features extracted from the *Wikidata*[3] knowledge base. In particular, each pair predicate-object is represented as a binary feature which is 1 if

---

The dialogue datasets generated from *MovieLens 1M* and *MovieTweetings* datasets can be found at: `http://github.com/swapUniba/ConvRecSysDataset`.
The source code of the automatic procedure for generating conversational recommender systems datasets will be released when the paper will be accepted.

[3] http://www.wikidata.org

the knowledge base contains the triple (item, predicate, object), 0 otherwise. The considered predicates are *wdt:P57* (director), *wdt:P161* (cast member), *wdt:P136* (genre) [4]. The dialogue generation procedure is an iterative algorithm which is executed until all user preferences have been used. At each step of the dialog generation procedure, a top-n list of items composed by positive and negative items is generated by randomly choosing from positive and negative preferences of the given user u. Then, paths from the root of the decision tree to the *consistently* classified examples are exploited to generate a sequence of questions, randomly chosen according to a binomial distribution over the item features, to elicit user preferences. Depending from the percentage of positive items in the top-n, a "refine" step is triggered which extends the dialog with additional questions that lead to a list of suggestions which contains only positive items.

Table 1 shows a conversation generated by applying the designed procedure to the well-known *MovieLens 1M* recommender systems dataset. In the first part of the conversation, utterances with the aim of introducing the user are generated by exploiting the contextual information included in the dataset.

Agent	**Utterance**
user	Hey
bot	What is your name?
user	I am Sofia
bot	How old are you?
user	I am 25
bot	What is your occupation?
user	sales/marketing
bot	Which are your favourite movies?
user	I love Speed
bot	I suggest you Erin_Brockovich Witness Pocahontas Four_Weddings_and_a_Funeral Anna_and_the_King. Do you like them?
user	I hate them
bot	What are your favourite directors?
user	I like John_Waters
bot	What are your favourite actors?
user	I like Josef_Sommer Wade_Williams Marg_Helgenberger Jeroen_Krabb
bot	What genres do you like?
user	My favourite genres are teen_film romance_film biographical_film
bot	I suggest you Erin_Brockovich Witness Ever_After Simply_Irresistible Hairspray. Do you like them?
user	I like them
bot	I am glad that you like them

Table 1. Conversation generated from the *MovieLens 1M* dataset. To help reading, *Wikidata* ids have been replaced with the corrisponding entities.

In this work we have proposed an automatic procedure able to generate synthetic dialogue datasets starting from well-known datasets in the recommender system field. The presented procedure is completely generic and can be applied on any dataset containing binary user preferences and whose items have a corresponding identifier in the Linked Open Data Cloud.

References

1. Mahmood, T., Ricci, F.: Improving recommender systems with adaptive conversational strategies. In: Proceedings of the 20th ACM conference on Hypertext and hypermedia. pp. 73–82. ACM (2009)

---

[4] The prefix wdt stands for http://www.wikidata.org/prop/direct/