Auto-generation of the customer questions and their ranking in e-commerce system

Yu Yu Dyulicheva
V I Vernadsky Crimean Federal University, Simferopol, Russia

E-mail: dyulicheva@gmail.com

Abstract. The paper proposes the framework for auto-generation of the customer questions to help customers make choice of the best products based on their needs. The task of the optimal search organisation from millions of products is crucial for e-commerce systems. We propose the approach based on the following stages: the web-scraping stage of products reviews sites, pre-processing stage to detect key-phrases based on TextRank algorithm and POS-tagging, the subjective probabilities estimation stage to detect questions estimations and their ranking based on TextRank algorithm and Bayesian rule.

1. Introduction and related works
There are millions of products and their titles, descriptions, customer reviews, customer questions and answers etc. on the e-commerce sites. The analysis of such big data can help to create a system that better understands its customers. Many problems arise in e-commerce systems such as auto-generation titles and descriptions of products, the intelligent assistant development, the products recommender systems development, the auto-generation of new products images and their texture recognition, research of multiple customers’ feedback signals such as order rate and click rates etc. The solving of these problems aimed at the customer personalization in e-commerce system development is based on the usage of the machine learning methods.

The problem of auto-generating products titles as a summary from their descriptions is considered in the paper [1]. Many methods are used such as n-gram counts, idf, removing noisy slot-value pairs and recombination of n-grams based on heuristic stack-based search algorithm for titles auto-generations. The quality estimation of the generating product titles is considered as classification problem and is solved based on random forest and Bidirectional Long Short-Term Memory model. The system for auto-generation answers on customer questions based on products attributes and customer reviews is developed based on PAAG algorithm that is composed of three components: a question-aware review representation module, a key-value memory network encoding attributes, and a recurrent neural network as a sequence generator [2]. The service for instant responses for product questions based on the reviews sites fragments, QA collections and existing pairs of questions-answers from amazon.com and the ranking of reviews based on positional language model is developed in [3]. The system for ranking products on positive, negative and neutral based on sentiment analysis of the product reviews and Naive Bayes approach is proposed in [4]. The development of the strategies of smart products search are very important for effectiveness of e-commerce systems based on the optimal customers choice. The approach of multi-lingual title generation for e-commerce browse pages based on sequence-to-sequence models are considered in the paper [5]. The web-pages of products description usually contain many
information and make it difficult to find data that is relevant to the customer requests. The framework development based on cascade of the ontologies models taking into account the learned distributional semantics and the semantics imposed by structured resource is described in [6]. The similarity-based ranking model, annotation-based classification model and ensemble matching model are considered as the question-answer matching model in the framework. The interesting research of the title, category, tags, description auto-generation based on the detection of the contextual relationship on the product images from online market site Etsy data with the usage of the combination of a convolutional neural network and recurrent neural network is completed in [7].

The task of developing intelligent services and intelligent assistants to help make the best product choice for customers’ needs is becoming one of the top priorities in e-commerce systems. Most modern online stores are equipped with the simplest product filters such as by color, by brand, by price etc. However, such simple filters are not able to satisfy all the customers wishes. Besides, many customers are faced with the problem of the questions formulations for online intelligent assistance aimed at the best product choice in the e-commerce system. In the paper, we propose the approach to the questions auto-generation and their ranking based on the TextRank algorithm and Bayesian rules.

2. The hybrid approach based on Bayesian rule and TextRank algorithm for the questions ranking

Let $H_1, ..., H_k$ be the categories of the products. The questions $Q_{i1}, ..., Q_{im_c}$ are assigned to each of $i$-th category of products. Let’s consider some tasks:

- how to auto-detect the names of product categories;
- how to select keywords that describe the products category as the basis for questions auto-generation;
- how to auto-generate subjective probabilities and how to range questions;
- how to calculate the best category of products based on the 3-5 auto-generating questions.

Consider the main stages of the hybrid approach based on Bayesian rule and TextRank algorithm proposed in the paper.

Web-scraping stage. At this stage, database is created that contains positive and negative reviews about the products of detected categories from scraping of the reviews sites (for example, t3.com, cnet.com, rthings.com) based on attributes such as thumb_up, thumb_down, good, bad, reasons to buy, reasons to avoid etc., titles and short descriptions of products from amazon.com. The names of categories are scraped from the menu items on review products sites.

Pre-processing stage. At this stage, we use TextRank algorithm for keywords detection and nltk python-library for stop words removing, sentence and words tokenization and POS-tagging of the products positive and negative reviews based on some templates such as JJ+NN, RB+NN, RBR+NN etc. and templates for the key-phrases generation that are the basis for questions auto-generation process.

For each keyword, we form vocabulary from its neighboring text units. Such vocabulary allows to identify semantically related questions.

Subjective probabilities estimation stage. At this stage, we need the approach to auto-calculate the subjective probabilities such as $P(Q_{ij}|H_i)$ - the probability that keyword defining the question $Q_{ij}$ is met in the positive reviews of products from category $H_i$ and $P(Q_{ij}|\overline{H_i})$ - the probability that keyword defining the question $Q_{ij}$ is met in negative reviews of products from category $H_i$. The estimation of the probabilities $P(Q_{ij}|H_i)$ and $P(Q_{ij}|\overline{H_i})$ are based on TextRank algorithm.

For two text units $V_i$ and $V_j$ with weight $w_{ij}$ the chances of two text units co-occurrence in a text window with fixed size [8] can be calculated as

$$W(V_i) = (1 - d) + d \sum_{V_j \in \text{inner}(V_i)} \frac{w_{ij}}{\sum_{V_k \in \text{out}(V_j)} w_{jk}} W(V_j)$$
Then, the conditional probabilities $P(Q_{ij}|H_i)$ are evaluated with the values

$$W(V_i)_{\max_{v \in PR(W(V_i)))}}$$

and the conditional probabilities $P(Q_{ij}|\overline{H}_i)$ are evaluated with the values

$$W(V_i)_{r \max_{v \in NR(W(V_i)))}}$$

where $PR, NR$ – the sets of the keywords from positive and negative reviews respectively that are scrapped from reviews sites, $r$ – empirically determined coefficient no more than 0.5.

Estimation and ranking the auto-generated questions stage. The estimation of questions are calculated as in Neylor expert systems [9]

$$C(Q_{ij}) = \sum_{j=1}^{m_i} \left[ \frac{P(Q_{ij}|H_i)P(H_i)}{P(Q_{ij}|H_i)P(H_i) + P(Q_{ij}|\overline{H}_i)P(\overline{H}_i)} - \frac{1 - P(Q_{ij}|H_i)P(H_i)}{1 - P(Q_{ij}|H_i)P(H_i) - P(Q_{ij}|\overline{H}_i)P(\overline{H}_i)} \right]$$

Based on the question estimations we get four questions with maximum estimations, ranging them on descending and present them to customers for choice. Detection of the best category stage. The category of products with maximum value of probability calculated based on Bayesian rule is the best category of products. In the initial stage of our hybrid approach the probabilities of categories are detected based on Dirichlet distribution and recalculated based on Bayesian rule

$$P(H_i|Q_{ij}) = \frac{P(Q_{ij}|H_i)P(H_i)}{\sum_{j=1}^{m_i} P(Q_{ij}|H_i)P(H_i)}$$

3. Experiments and results

We concluded experiments on the headphones categories from amazon.com. As the result of scraping stage we had 8 categories of products based on reviews product sites such as Best Noise Cancelling, Best gaming headphones, Best Wireless Bluetooth Earbuds, Best SoundSport Free Truly Wireless, Best headphones for women and girls, Best headphones for kids, Best headphones for teleconference and video calling and Best DJ headphones.

On the preprocessing stage, we got keywords and their values from positive and negative headphones reviews based on the TextRank algorithm. For example, such keywords were auto-generated for the category named “the best soundsport free truly wireless” as {'sport': 2.93448016; 'fit': 2.710578805856; 'sound': 2.22954541; 'battery life': 2.20367196; 'headset': 1.94549744; 'design': 1.825857394; 'earphones': 1.610094 etc.} based on positive reviews and {'design': 2.08901575; 'battery life': 1.8445688036; 'quality': 1.8210775933, 'noise': 1.460239771 etc.} based on negative reviews. On this stage, we also got two descriptive vocabularies for each keywords based on positive and negative reviews. For example, for the keyword “sound” the framework was generated such positive descriptive vocabulary {'crisp', 'great', 'amazing', 'decent etc.'}. The auto-generated questions before removing semantically close questions are shown on screenshot 1.

![Figure 1](image-url) Figure 1. The result of questions auto-generation stage.
We constructed the templates of questions such as template_questions = [{'advise, recommend': 'What headphones can you VB with key-phrase?', 'prefer, like': 'I VB key-phrase. Do you have?' etc}] and substitutions based on regular expressions. Than we calculated estimations of the probabilities $P(Q_{ij}|H_i)$ and $P(Q_{ij}|\bar{H}_i)$ for the ranking questions with and presented four questions with maximum values of estimations to the customer as shown on the screenshot 2.

![Screenshot of four best questions and three best products.](image)

**Figure 2.** Detection of four best questions and three best products.

4. Conclusion
In the paper the system aimed at helping the customer to formulate questions that allowed make the best choice of products category based on Bayesian rule and TextRank algorithm was realized.

In the future, we will plan to improve our system based on entity recognition of manually entered customer questions and the best product description based on the summary auto-generation from the brief description of products and verdicts from reviews sites.

References
[1] De Souza J G C, Kozielski M, Mathur P, Chang E, Guerini M, Negri M, Turchi M, Matusov E 2018 Generating e-commerce product titles and predicting their quality *Proceedings pf the 11th International Natural Language Generation Conference* pp 233-43
[2] Gao S, Ren Z, Zhao Y 2019 Product-Aware Answer generation in e-commerce question-answering *The Twelfth ACM International Conference on Web Search and Data Mining* ACM New York NY USA
[3] Qian Yu Lam W Wang Z 2018 Responding e-commerce product questions via exploiting QA collections and reviews *Proceedings of the 27th International Conference on Computational Linguistics* pp 2192-203
[4] Jayasree J, Mathi M Sri A 2017 Product ranking system in e-commerce website for validation using sentimental analysis *International Journal of Science & Engineering* 3 (9) 176-86
[5] Mathur P, Ueffing N, Leusch G 2018 Multi-lingual neural generation for e-commerce browse pages *Proceedings of NAACL-HLT* New Orleans pp 162-9
[6] Kulkarni A Mehta K Garg Sh 2019 ProductQnA: Answering user questions on e-commerce product pages *Proceedings of The 2019 World Wide Web Conference* pp 354-60
[7] Yashima T, Okazaki O Inui K 2016 Yamaguchi K Learning to describe e-commerce images from
noisy online data 13th Asian Conference on Computer Vision part V Springer International Publishing

[8] Zuo X Zhang S Xia J 2017 The enhancement of TextRank algorithm by using word2vec and its application on topic extraction Journal of Physics: Conference Series 887

[9] Chernoruckij I G 2005 The methods of decision making (S-Peterburg)