Using Textual Summaries to Describe a Set of Products

KITTIPITCH KUPTAVANICH, University of Aberdeen, UK

When customers are faced with the task of making a purchase in an unfamiliar product domain, it might be useful to provide them with an overview of the product set to help them understand what they can expect. In this paper we present and evaluate a method to summarise sets of products in natural language, focusing on the price range, common product features across the set, and product features that impact on price. In our study, participants reported that they found our summaries useful, but we found no evidence that the summaries influenced the selections made by participants.

CCS Concepts: • **Computing methodologies** → **Natural language generation**; • **Information systems** → *Recommender systems*;

Additional Key Words and Phrases: NLG, Recommender System

ACM Reference Format:
Kittipitch Kuptavanich. 2018. Using Textual Summaries to Describe a Set of Products. 1, 1 (July 2018), 8 pages. https://doi.org/0000001.0000001

1 INTRODUCTION

Tanner and Raymond [20] observe that, when consumers are faced with a set of products to choose from, most would not be interested in examining them exhaustively. Instead, they start with *"information searching"*. During information searching, customers acquire a basic overview of the product set as a whole [20]: for instance, how many subtypes there are in the set, what features to expect in a typical product of this type, what is the price range, and so on. Based on this information, the consumer can then develop *"evaluation criteria"* to help narrowing down choices.

We believe that Tanner and Raymond's ideas are relevant to Recommender Systems (RS) in a number of ways. An overview of a set of products can enhance an RS, by increasing trust, effectiveness, persuasiveness, efficiency and satisfaction [21] when consumers engage in Tanner/Raymond-style information searching upon the set. Additionally, where an RS has insufficient information to confidently recommend one item (e.g., during a "cold start") [18], then a Recommender System may opt to present a larger set of items, necessitating Tanner/Raymond-style information searching.

My PhD project explores the hypothesis that a consumer's decision making can be aided by an approach inspired by Shneiderman's Visual Information Seeking mantra [19]. This mantra is often summarised by the slogan, "*Overview first*, zoom and filter, then details-on-demand". We focus here on the first part of Shneiderman's slogan (italicised), namely that it is beneficial for a reader to be exposed to an overview before diving into specifics.

Textual overviews of large sets of consumer products are not a new idea, of course. However, we observe that such overviews are written by hand [2, 24]. While static manual summaries are typically provided for only top level categories, dynamic computer generated summaries can be provided for any set of products, for example those filtered by users with their criteria of interest. Our research hypotheses is that useful overviews of consumer product sets can be generated automatically by computer, using Natural Language Generation (NLG) techniques. If confirmed, this would be a potentially important finding for both the Recommender Systems and the Information Retrieval community.

Author's address: Kittipitch Kuptavanich, University of Aberdeen, UK, kittipitch@gmail.com.

2018. Manuscript submitted to ACM

To illustrate my work so far, I will present description of our automatically generated summaries inspired by handwritten summaries, followed by an evaluation with human users. Section 2 summarises relevant work in NLG and automated summarisation. Section 3 describes our automatically generated summaries based on our observation over handwritten summaries. Section 4 presents an experiment in which we tried to assess whether the summaries generated by our algorithm helps users "understand" the information in two product sets, and whether it helps them make a well-founded choice. Section 5 discusses how our findings point the way towards even more useful computer-generated product summaries.

2 RELATED WORK

2.1 Product Comparison Interfaces

Many websites support product comparison, mostly in a tabular format, where users can select 2 or more products and features to compare side-by-side [4, 17]. Consumers are often able to apply feature based filters to narrow down the set of products [10, 12, 22]. Product information is often presented as a table of specifications with no accompanying textual summary of the items presented in the table.

2.2 Natural Language Generation

Within Computational Linguistics, the automatic generation of text from non-textual input is addressed in the research area of Natural Language Generation (NLG) [5, 15].

One area of NLG that is potentially relevant here is Referring Expressions Generation, where the aim is to identify a referent for a hearer (i.e., so the reader knows what it is). As in our case, the referent may be a set, for instance as when an NLG system generates "the blue sofas" to enable a reader to know what sofas the writer has in mind [23]. Unlike most previous work, however, our aim is not to allow the reader to know *which consumer items* the system is talking about: the purpose, rather, is to give the reader insight in the broad composition of the set (e.g., so s/he knows what the main commonalities and difference across the set are).[7]

Given what we observed about the prevalence of quantified statements in product set surveys (section 3), another potentially area of NLG is where NLG algorithms extract trends and patterns from data. Perhaps the most sophisticated example is Narratives for Tableau [9], a commercial product that generates text from analyzing data associated with user-selected areas of a chart made with Tableau. The extension produces description of the data such as *"Sales and profit ratio moved in opposite directions from January 2011 to December 2014."* However, the extension focuses on time-series data, and it is difficult to see how Tableau's techniques could be used for giving insight into a static set of product. A similar example is Automatic Statistician [1], which also does time-series data to text generation using statistical methods. To our knowledge up to this point, there is no NLG system that describes set of items.

2.3 Explanation of Recommender System

Item description accompanying recommended set can be designed to benefit readers with various aims in mind. [21] classified them as Transparency (Tra) - Explain how the system works, Scrutability (Scr) - Allow users to tell the system it is wrong, Trust - Increase users confidence in the system, Effectiveness (Efk) - Help users make good decisions, Persuasiveness (Pers) - Convince users to try or buy, Efficiency (Efc) - Help users make decisions faster and Satisfaction (Sat) - Increase the ease of usability or enjoyment. A dynamically generated description of recommended items, when used as an explanation, could potentially be an enhancement to a Recommender System in multiple aspects.

3 AUTOMATIC SUMMARISATION OF PRODUCT SETS

Many websites contain hand-written reviews of product sets.[2–4, 17, 22, 24]. Here we list the lessons we observed that informed the algorithm we used to generate our first summary. As expected, most reviews started with an *introductory paragraph* that surveyed the product set as a whole and sketched the shape of the price curve of the product set. Subsequently, many reviews contained sentences that *quantify* over the product set, using patterns like *"Most [products] are/have..."*, *"Many [products] have..."*. Reviews also tended to say which features they should *pay attention to*.

As a result of our observation, we implemented an NLG system using simple template-based approach with jinja2 template engine [16]. From the 4 stages of data-to-text NLG architecture [13], the majority of our work belongs to the content selection process of the *Document Planning* stage. That generated summaries consisted of 3 parts namely, an introductory paragraph, followed by a collective description of products, and important feature highlights each in their own paragraph as shown below.

> *For 32 inch TVs, the price of most products (340 out of 363 models) falls in the range of 70-580 pounds with a median price of about 255 pounds.*
>
> *Most 32 inch TVs have following features: 16:9 aspect ratio, LED backlight, LCD display technology, HDMI, Flat panel design, analogue TV tuner, and digital TV tuner*
>
> *The features that have a strong impact on the price of 32 inch TVs are: number of hdmi inputs, release year, brightness, resolution, hd ready 1080p (full hd), smart TV, and annual energy consumption*

Introductory Paragraph: Shape of the price curve. In this part of the summary, we simplified the task by focusing on providing information describing the shape of the price curve of the set of product only. The price shape of the product are reported as a range of price, without the outliers, rounded to the nearest 5 (the outliers were identified using median absolute deviation: MAD).

Collective description of products: Common Features. In this part of the summary, we chose to report the most common features across the set of products. From the set of products we identified the 7 top-most common features then reported it.

Highlighting Important Features. Since the items we are trying to report are consumer products, we theorised that the features that display strong effects on the price are the important ones that the consumers should focus on.

So for each features, we find average price for each subset. For example, if an interested feature of a TV is 'resolution', then the subgroup are 720p, 1080p, and 4K. We then find the average price of each subgroup. Then we again find the SD of the 3 average prices. The higher the SD, the greater effect on the price that feature has. We then ranked the top most 7 features using this comparison and generated a report.

4 EVALUATION EXPERIMENT

NLG algorithms have traditionally been evaluated in a number of ways [14]. Given that our present aim is to produce texts that are useful to a reader (rather than texts that mimic a speaker), metric-based evaluations, such as BLEU for instance [11], are not very suitable. It therefore seemed to us that the most apt methods in our case are evaluation by means of human judgment and task-based evaluation. We decided to conduct a simple version of each of these evaluation methods, to see what they might teach us about our algorithm.

We wanted to find out whether automatically generated summaries of large sets of products can help customers make an informed decision about what product to buy; to do this, we focused on summaries generated by the above

algorithm, which focuses on listing common product features and on listing features that have a strong effect on price. We wanted to know two things in particular: first, we wanted to find out how useful readers *believe* our summaries to be for selecting the products of interest to them; second, we wanted to make a first attempt at finding out whether our summaries actually *helped* participants to quickly identify those products that they are interested in.

To answer the first question (about participants' subjective appreciation of the summaries) we asked participants to answer four Likert-style questions that address the *perceived* usefulness of the summaries (which were mapped to different aims of explanation); to answer the second question (about the actual usefulness of the summaries) we asked participants to make a quick choice ("speeded choice") from among all the products in the set after reading our summaries, and we compared these speeded choices with the choices that they would later make at their leisure ("gold-standard choice"); the smaller the difference between speeded choice (facilitated by our summaries) and gold-standard choice (reflecting participants' real preference), the more useful we considered the summaries to be. Our experiment thus consisted of Laboratory Human Rating and laboratory task-based evaluation. This idea will be explained and discussed in the following sections.

4.1 Method

4.1.1 Materials. We used two different product information databases:

- database of TVs containing 363 products (rows) with 100 features (columns)
- database of photo cameras containing 610 products (rows) with 71 features (columns)

Both databases contained data scraped from [12] during June 2017. They were presented as spreadsheets in MS Excel format.

We chose large databases to make the task of "understanding" their content and selecting the most interesting items in them particularly challenging.

For the baseline group, we used an ultra-short summary that was designed to be truthful without being particularly helpful, as shown below.

Example Summary in Baseline Condition:

For DSLR Cameras, the price of most products (550 out of 610 models) falls in the range of 10-1850 pounds with a median price of about 525 pounds.

For the experimental group, we used the summary produced by our algorithm.

Example Summary in Full Summary Condition:

For 32 inch TVs, the price of most products (340 out of 363 models) falls in the range of 70-580 pounds with a median price of about 255 pounds. Most 32 inch TVs have following features: (followed by 7 features). The features that have a strong impact on the price of 32 inch TVs are: (followed by 7 features)

4.1.2 Participants. Participants were 16 graduate students in Computing Science and Chemistry Department of (ANOMYMISED) recruited through the departments' internal student mailing lists.

4.1.3 Design and Procedure. In total, there were thus 4 conditions:

- Condition A: Camera + Baseline Summary, TV + Full Summary
- Condition B: TV + Full Summary, Camera + Baseline Summary
- Condition C: Camera + Full Summary, TV + Baseline Summary
- Condition D: TV + Baseline Summary, Camera + Full Summary

To find out about participants' subjective appreciation, we asked each participant a number of Likert-style questions about the usefulness of the summaries that they had seen. We used 5 Likert values, from 1 (I do not agree at all) to 5 (I fully agree).

- "Does the summary help you get a rough picture of the products in this category?" [Q1] – Trust
- "Does the summary help you select the products from the table faster?" [Q2] – Efficiency
- "Does the summary help you select the products from the table with more confidence?" [Q3] – Effectiveness
- "Do you find the summary useful?" [Q4] – Satisfaction

To find out about the effectiveness of the summaries, We asked each participant to list their top-5 products (e.g., their top-5 TVs). Participants were explicitly told that the order of the items in their top-5 list did not matter. Crucially, we asked them to twice produce such a list: once after they had looked at the database for only 5 minutes (Speeded Choice; we call the resulting list of products the *Speeded Set*), and once after they had studied the database extensively (gold-standard Choice; we call the resulting list of products the *gold-standard Set*). Our expectation was that the Speeded and gold-standard Sets produced by participants in the Full Summary condition would be more similar to each other than those of participants in the Baseline Summary condition, because the Full Summary had given them a head start in understanding the database of products.

Each participant was asked to do 2 categories of products. During a pilot experiment, we had observed that some participants had used the 'Filter' feature in MS Excel, which had had a very substantial effect on the time used. For the experiment itself we therefore included a question asking participants whether they had used the feature during the experiment. We also asked participants what information they thought they would want to see added to the summary. Finally, to ensure that participants would take the experiment seriously, we offered 50 pounds reward for the "best" list of products and 20 pounds for the most valuable feed back. – To summarise the procedure:

(1) The participants read the summaries for a minute.
(2) Working with a product database spreadsheet, the participants were given 5 minutes to write their Speeded List on a piece of paper.
(3) With another database spreadsheet, the participants were given another 10 minutes to write their gold-standard Set on another piece of paper.
(4) Afterward the participants answered a questionnaire containing the Likert questions and then key in their preferred product lists on Google Form.
(5) Then we repeated the procedure for another product category.

4.1.4 Hypotheses. Our hypotheses were:

Hypothesis 1 [H1]: Likert scores are better for Full Summaries than for Baseline Summaries

Hypothesis 2 [H2]: The similarity between Speeded Set and gold-standard Set is greater for participants in the Full Summary conditions than for participants in the Baseline Summary conditions.

For Hypothesis 2, since the sets of products were unordered, we computed similarity between sets of products using the Dice score, a well-known formula for assessing the similarity of sets:

$$Dice(Speeded, GoldStandard) = \frac{2 \times |Speeded \cap GoldStandard|}{|Speeded| + |GoldStandard|} \quad (1)$$

Here *Speeded* is the set of attributes expressed in the description produced by a human author and *GoldStandard* is the set of attributes expressed in the Logical Form generated by an algorithm. Dice yields a value between 0 (no agreement) and 1 (perfect agreement).

4.2 Results

4.2.1 Subjective appreciation. For the Likert part of the experiment, in which we addressed participants' subjective appreciation of the summaries, we found that the participants liked the full summaries better than the baseline summaries in all four respects (i.e., regarding all 4 questions), with statistical significance at p = 0.05 both before and after Bonferroni Correction (Table 1).

Table 1. Likert scores of the 4 Questions asked

	Q1	Q2	Q3	Q4
N [baseline]	16	16	16	16
N [exp]	16	16	16	16
mean [baseline]	2.50	2.44	2.06	2.69
mean [exp]	4.00	3.69	3.69	3.88
SD [baseline]	1.26	1.26	1.00	1.14
SD [exp]	0.82	1.01	0.95	0.81
p-value (raw)	0.0004	0.0043	0.0001	0.0019
p-value (Bonferroni Correction)	0.0016	0.0172	0.0004	0.0076

- [baseline] denotes groups with Baseline Summary
- [exp] denotes groups with Full Summary

4.2.2 Task performance. There was no significant difference, however, between the Dice coefficient from the Baseline Summary (0.1375) and the Full Summary Group (0.1250) (Table 2); in fact, the Baseline Summaries performed marginally better (not statistically significant) than the Full Summaries.

Table 2. Similarity of sets produced in the 2 steps (Dice)

	Dice coefficient
N [baseline summary]	16
N [full summary]	16
mean [baseline]	0.14
mean [full summary]	0.13
SD [baseline]	0.17
SD [full summary]	0.26
p-value	0.8749

Here, Hypothesis 1 have been confirmed by the result while Hypothesis 2 is inconclusive.

5 DISCUSSION AND FUTURE WORK

Results relating to Hypothesis 1 indicate that participants regarded the summaries generated by our algorithm as more useful than Baseline Summaries from the point of view of understanding the product database and selecting a product

rapidly and confidently. It seems plausible that the list of common features and price influential features from the summary helped participants to know which columns of the database they should focus their attention on.

The fact that we were unable to confirm Hypothesis 2 raises interesting questions. We wondered whether a more sophisticated measure of set similarity might lead to a different result, but it turned out that a metric based on cosine similarity (which acknowledges that two products might be different yet share most of their features) did not confirm Hypothesis 2 either.

We considered several possible explanations for the mismatch between subjective appreciation and task performance. Based on the psychology literature on cognitive dissonance reduction [6], one possibility is that participants were reluctant to change the set of product that they had chosen as their Speeded Set. If this was true, one might expect to see gold-standard Sets that were highly similar to the Speeded Sets in both the Baseline Summary condition and the Full Summary condition. However, the low Dice scores (0.14 and 0.13) in Table 2 show that this was not the case.

It is possible that our full summaries appeared useful to participants but that in reality, they were not. Asymmetries of this kind, between perceived and actual usefulness, are surprisingly common; an example is [8] where doctors reported a preference for graphical over textual information presentation although their task performance with textual information was better.

In our view, a more likely explanation is that the setup of our experiment did not do full justice to the idea of Hypothesis 2. In particular, the 10 minutes offered to subjects in the gold-standard choice condition may have been too short. Thus, what we had meant to be a gold standard may not have been an accurate reflection of participants' real preferences (i.e., not a genuine gold standard). In reality, people would often spend a lot more time buying expensive products with their budget limit in mind.

In the remainder of this doctoral research, we will explore these issues further.

Following up on the above experiment, and making use of a more detailed study of our corpus of human-written summaries, we have modified our algorithm in a number of ways. First and foremost, we have amplified our summaries to contain a comparison between the products in the target set (e.g., 32-inch TVs) and the products in a natural superset (e.g., TVs). Second, statistical analysis of the corpus has been used to select important features to mention in the summary, and what sentence patterns are employed to talk about them; the patterns involved quantify how frequently a given feature occurs in the target set, for instance "Most TVs in this category have an HDMI port", "Only a few TVs have 4K resolution". Lastly, a short sentence describing how each importance feature influences prices is included, for example, "TVs with smart features are more expensive in average".

In this paper, we have focused on the ability of summaries to provide insight in the content of a set of products. In future research, we want to explore a closely related idea, namely the possibility of using automatically generated summaries of a set of products to *explain why* a certain recommendation (i.e., a recommendation for one or more members of the set) is made. The project will address 2 scenarios of set description, which are: descriptions of a predefined set of objects (from databases) in general, and description of a set resulting from a search or a Q/A process. We also plan to extend the algorithm to generate the list of important features by building and analyzing corpora on recommendations over different product categories available online. When employed in this manner, summaries might be able to boost users' trust in the recommendation and the Recommender System itself.

REFERENCES

[1] automaticstatistician.com. 2017. Automatic Statistician. (2017). https://www.automaticstatistician.com/ [Online; accessed 25-August-2017].

[2] ConsumerReports. 2017. Consumer Reports: Product Reviews and Ratings. (2017). https://www.consumerreports.org/cro/index.htm [Online; accessed 8-August-2017].

[3] DigitalPhotographyReview. 2017. Digital Photography Review. (2017). https://www.dpreview.com [Online; accessed 8-August-2017].

[4] GadgetsNow. 2017. Technology News, Latest & Popular Gadgets Reviews, Specifications, Prices, Mobile Comparison, Technology Videos & Photos | Gadgets Now. (2017). https://www.gadgetsnow.com [Online; accessed 8-August-2017].

[5] Albert Gatt and Emiel Krahmer. 2017. Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation. *arXiv preprint* c (mar 2017), 1–111. arXiv:1703.09902

[6] Johanna M. Jarcho, Elliot T. Berkman, and Matthew D. Lieberman. 2011. The neural basis of rationalization: cognitive dissonance reduction during decision-making. *Social Cognitive and Affective Neuroscience* 6, 4 (2011), 460–467. https://doi.org/10.1093/scan/nsq054

[7] Roman Kutlak. 2014. *Generation of Referring Expressions for an Unknown Audience.* Ph.D. Dissertation. University of Aberdeen.

[8] Anna S. Law, Yvonne Freer, Jim Hunter, Robert H. Logie, Neil Mcintosh, and John Quinn. 2005. A Comparison of Graphical and Textual Presentations of Time Series Data to Support Medical Decision Making in the Neonatal Intensive Care Unit. *Journal of Clinical Monitoring and Computing* 19, 3 (01 Jun 2005), 183–194. https://doi.org/10.1007/s10877-005-0879-3

[9] narrativescience.com. 2017. Narratives for Tableau - Natural Language Generation for Tableau. (2017). https://narrativescience.com/Partners/Business-Intelligence/Tableau [Online; accessed 25-August-2017].

[10] Opodo. 2017. Book cheap holidays: flights, hotels and car hire - Opodo. (2017). https://opodo.co.uk [Online; accessed 8-August-2017].

[11] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 311–318. https://doi.org/10.3115/1073083.1073135

[12] PriceSpy. 2017. PriceSpy - Compare prices and do more with your money! (2017). https://pricespy.co.uk [Online; accessed 8-August-2017].

[13] Ehud Reiter. 2007. An Architecture for Data-to-Text systems. *ENLG 2007 - Eleventh European Workshop on Natural Language Generation* (2007), 97–104. https://doi.org/10.1017/CBO9781107415324.004 arXiv:arXiv:1011.1669v3

[14] Ehud Reiter. 2017. Types of NLG Evaluation: Which is Right for Me? (2017). https://ehudreiter.com/2017/01/19/types-of-nlg-evaluation/ [Online; accessed 20-November-2017].

[15] Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems.* Cambridge University Press, New York, NY, USA.

[16] Armin Ronacher. 2018. Jinja2. (2018). http://jinja.pocoo.org/ [Online; accessed 20-Apr-2018].

[17] SaveonLaptops. 2017. Save on Laptops | Cheap Laptops | UKs Best Laptop Deals. (2017). www.saveonlaptops.co.uk [Online; accessed 8-August-2017].

[18] Andrew I. Schein, Alexandrin Popescul, Lyle H. Ungar, and David M. Pennock. 2002. Methods and Metrics for Cold-start Recommendations. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '02)*. ACM, New York, NY, USA, 253–260. https://doi.org/10.1145/564376.564421

[19] B. Shneiderman. 1996. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*. 336–343. https://doi.org/10.1109/VL.1996.545307

[20] John F Tanner and Mary Anne Raymond. 2011. *Principles of marketing.* Flat World Knowledge, Irvington, N.Y.

[21] Nava Tintarev and Judith Masthoff. 2007. A survey of explanations in recommender systems. *Proceedings - International Conference on Data Engineering* (2007), 801–810. https://doi.org/10.1109/ICDEW.2007.4401070

[22] uSwitch. 2017. Energy Comparison of Gas & Electricity | Broadband Deals & Mobile Phones | uSwitch.com. (2017). https://www.uswitch.com [Online; accessed 8-August-2017].

[23] Kees Van Deemter. 2002. Generating Referring Expressions: Boolean Extensions of the Incremental Algorithm. *Computational Linguistics* 28, 1 (2002), 37–52. https://doi.org/10.1162/089120102317341765

[24] Which? 2017. Reviews and expert advice from Which? (2017). http://www.which.co.uk [Online; accessed 8-August-2017].