Generating Summaries of Sets of Consumer Products: Learning from Experiments

Kittipitch Kuptavanich
Department of Computing Science
University of Aberdeen
Aberdeen, UK
kittipitch.kuptavanich@abdn.ac.uk

Ehud Reiter
Department of Computing Science
University of Aberdeen
Aberdeen, UK
e.reiter@abdn.ac.uk

Kees Van Deemter
Utrecht University
Utrecht, NL
k.vandeemter@uu.nl

Advaith Siddharthan
Knowledge Media Institute
The Open University
Milton Keynes, UK
advaith.siddharthan@open.ac.uk

Abstract

We explored the task of creating a textual summary describing a large set of objects characterised by a small number of features using an e-commerce dataset. When a set of consumer products is large and varied, it can be difficult for a consumer to understand how the products in the set differ; consequently, it can be challenging to choose the most suitable product from the set. To assist consumers, we generated high-level summaries of product sets. Two generation algorithms are presented, discussed, and evaluated with human users. Our evaluation results suggest a positive contribution to consumers' understanding of the domain.

1 Introduction

When presented with a large amount of data in tabular form, an additional textual summary could aid a reader's comprehension of the otherwise overwhelming information at hand. The task of automatically creating a summary from numerical data is an ongoing research area within Natural Language Generation (NLG). We explored this task in the context of generating a textual summary describing a large set of objects [products] from a large database, where each object is characterised by several product features.

Product set overviews can be written by hand if the category is known beforehand. For example, manually written product reviews often start with an overview paragraph that discusses a wider set of products of which the product is a member. However, when a consumer searches for products with keywords or through filters (e.g. on an e-commerce website), an overview of the returned set of search results would have to be automatically generated.

In this paper, we test the hypothesis that automatically generated textual summaries can be of benefit to customers. This can be seen as a specific instance of Shneiderman's Visual Information Seeking mantra (Shneiderman, 1996) of "Overview first, zoom and filter, then details-on-demand". One of the main ideas presented there is that it is beneficial for a reader to be exposed to an overview of the information before diving into specific details of interest.

There have been related NLG research about sets of objects, although with different goals or focuses. For example, to refer to or identify a set of objects within a larger set (Van Deemter, 2002), to perform a data-to-text analysis of tabularized data by records[1], to generate a page title for set items with shared characteristics from existing metadata (Mathur et al., 2017), or to address the issue of missing data encountered in summarisation (Inglis et al., 2017). In contrast, our work explores summaries that describe commonalities and differences within a set in order to help a user make informed decisions in selecting an object from the set. Our work focuses particularly on Content Determination step in the NLG pipe-line (Reiter and Dale, 2000), including selecting features and values to be presented.

2 Analysis of Hand-written Reviews

To inform our algorithms, we manually analysed 30 hand-written reviews gathered with the search term "best TV review" on Google. We used

[1] www.ax-semantics.com

Proceedings of The 11th International Natural Language Generation Conference, pages 403–407,
Tilburg, The Netherlands, November 5-8, 2018.

the 30 top ranked pages which contained a list of TVs (not just one single product). We then defined a per-clause tagging scheme to identify aspects that could be generated from product specifications and to systematically observe how reviewers described sets of products. In our scheme, a clause could have multiple tags. There was one annotator involved in the tagging (the first author). Our finding are summarised below.

Feature Selection: We analysed how often each product feature gets mentioned in the reviews. We found, as shown in Table 1, besides the price, the most frequent features (in descending order) are screen size, resolution, smart/internet feature, brand, backlight technology, ports, and contrast.

Feature	Frequency (%)
Screen Size	73
Resolution	60
Smart/Internet	43
Brand	40
Backlight Technology	30
Connectivity (Ports)	30
Contrast	30

Table 1: Frequency Count of Features in Reviews

Price Description: The product price in the reviews are typically mentioned only vaguely, using terms like "desirable price", "cheap", "expensive" or "premium". The description is vague even when numbers are involved e.g. "around £300". But when a crisp description is used it is more often found in the form of stating the starting point, e.g., "you can get a 1080p TV starting at £270" or the maximum e.g. "Discover the best 32 inch Smart TVs under £300 here".

Description of a Set of Items: Usually in a review, only a small number of sentences explicitly describe the set as a whole, for example "Most 32-inch TVs these days are labeled as *HD Ready*". When they do they uses quantifying words like "most". Numbers are described vaguely e.g. weight is mentioned as "light" or "response time" is either "fast" or "slow". Some features, for example the screen size, are mentioned both as exact numbers and vague description.

Price–Features Relationship: The relationship with price is used as a secondary justification to the features that the reviewers already think important, for example, "A TV with a 1920 × 1080 resolution [are] not even that much more expensive" or "good image quality and available smart features [...] carry a price premium."

Based on this analysis, we decided that our summaries should describe the shape of the price curve, the important features, and the effect of these features on price.

A large part of the reviews gathered included domain knowledge, for example, descriptions of technical terms and other insights. This part of the reviews clearly could not be produced from specification table. There were also mentions of features that can be, non trivially, derived from the table, e.g. picture quality (which can be based on columns like resolution and contrast).

3 The Algorithms

3.1 Alg1. Summarising a set of products

In our previous work (Kuptavanich, 2018), we presented an algorithm (called Alg1 here) to generate summaries consisting of (a) the shape of the price curve, (b) common features within the set and (c) features that influence price (Figure 1 gives an example of the generated text). The algorithm mainly used the influence of a feature on the product price to determine its importance.

> *As a result of your query for TVs, the price of most products in the result (432 out of 456 models) falls in the range of £275 - £1500 with a median price of about £630.*
> *Most TVs in this result have following features: 4K Ultra HD Display Resolution Max and FREE Shipping*
> *The features that have a strong impact on the price of TVs in this result are: Display Size, Image Aspect Ratio, Supported Content Service, Tuner Technology, Brand, Connectivity Technology, and Freeview Enabled.*

Figure 1: Alg1 Summary Example

3.2 Alg2. Dynamically summarising and contextualising a set of products

Alg1 only included content that could be generated from descriptions of items in a set being summarised. Following our analysis of the handwritten reviews, we adapted the algorithm. The resulting Alg2 allows for dynamic creation of sets through the use of feature filters and the contextualisation of these sets with respect to the unfiltered wider set as described below.

Shape of the Price Curve: Alg1 reports the median price and the price range of the set. [alg2] additionally compares the median price of the filtered set against the median price of the wider category. For instance, the first 3 lines of Figure 3

show a situation where the user has filtered the set of TVs to those that are 40–59 inches with 4K ultra high definition. The underlined portion is generated only by Alg2.

Description of Important Features: In the TV domain, the following features occurred most frequently: *display size, display resolution, smart/internet feature, support content service, brand, display technology, connectivity technology (ports) and HDR.* We therefore focussed on these features, but generated more detail about them than in Algo1. The description of each feature consisted of two parts. The first used quantifiers to describe the common values for the feature within the set. The second compared the median price of products with the said feature values against the median price of general products in this category and reported feature values that impacted on price (Figure 2).

> *Brand: Most products in the result are branded Samsung, Sony, LG Electronics, or Philips Brand. TVs with Samsung Brand are generally more expensive (£580 vs £450).*

Figure 2: Our Description of Important Features

(a) Quantifiers. Here we generated expressions such as *"Most products in the result are..."* If the values of a feature are continuous numbers e.g. weight, we report them in the same fashion as the price (i.e. range and median value). Otherwise, we use the quantifiers "most" (more than 50%), "a large proportion" (more than 25%) and "some" (more than 10%).

(b) Comparatives and Qualifiers. In the second part, we also use phrases such as "more expensive" , "less expensive" or "about the same price" (when the difference is less than 5%). If the difference falls between 5 - 10%, we qualify this using the word "slightly". This generates texts such as *"TVs with Smart-Internet Feature are generally slightly more expensive (£475 vs £450)."* (Figured 3, in the next section).

The processes from document structuring through realization was carried out through template/schemata approach (McKeown, 1985). Also, the tone of the discourse is primarily to provide factual product overview without trying to be persuasive. Both algorithm were implemented using the Jinja2[2]template engine.

[2]jinja.pocoo.org

4 Evaluation Experiment

Our previous work (Kuptavanich, 2018) revealed difficulties designing a suitable task that reflected real consumer behaviours in the task based experiments, but a promising result with human rating. We therefore decided only to focus on human rating evaluation[3]. The scenario of interest is where a consumer is searching for products on an e-commerce website. Our Laboratory Human Ratings Evaluation experiment had three goals. First, we wanted to find out whether the text summaries generated by Alg2 were preferred over those generated by Alg1, and also over the static introductory text provided on the e-commerce site. Second, we asked the participants to identify parts of the summaries that were useful, parts that were unnecessary, and what they want to see added. Third, we wanted to also find out what product features are important in the decision-making process.

4.1 Method

Materials: We scraped TV product data from Amazon UK[4] during May - June 2018 to obtain 1478 products. We used this database to generate the summaries using both Alg1 and Alg2. As our baseline, we used Amazon's static text provided on their TV browsing page. An excerpt is shown in Figure 4. The full text can be found on Amazon UK TVs[5] page.

We used two product search scenarios on Amazon UK, based on its search filters. Each scenario produced a different set of search results and thus generated different summaries for Alg1 and Alg2.

Participants: Participants were 18 graduate students in Computing Science and Chemistry Department of University of Aberdeen recruited through the departments' internal student mailing lists.

Design and Procedure: In total, there were 2 pre-determined product search scenarios:

- Scenario 1: 40 – 59 inch TVs with Ultra HD
- Scenario 2: TVs of any size that are smart TVs

First, summaries [amz], [alg1], and [alg2] were presented in random order. To ensure that participants engaged with the task, each participant was asked to select one product. Then they were asked

[3]https://ehudreiter.com/2017/01/19/types-of-nlg-evaluation

[4]www.amazon.co.uk

[5]www.amazon.co.uk/LED-Smart-4K-TVs/b?node=560864

Size (inches): 40 - 49, or 50 - 59
Price:
Keyword(s): 4K Ultra HD
–

From your search result, the price of most products falls in the region between £275 and £1500, with many models around £630 price point, making it more expensive than TVs in all categories combined on average (£630 vs £450). Below we list a number of features that might be of interest to most people.

Display Size: Most products in the result either have 55, or 49" Display Size. Some have 43". TVs with 55" Display Size are generally more expensive (£770 vs £450).

Display Resolution Max: Most products in the result have 4K Ultra HD Display Resolution Max. TVs with 4K Ultra HD Display Resolution Max are generally more expensive (£660 vs £450).

Feature: Most products in the result have Smart-Internet Feature. TVs with Smart-Internet Feature are generally slightly more expensive (£475 vs £450).

Supported Content Service: Most products in the result either have Netflix or YouTube Supported Content Service. Some have Amazon Video. TVs with Netflix Supported Content Service are generally about the same price (£440 vs £450).

Brand: Most products in the result are branded Samsung, Sony, LG Electronics, or Philips Brand. TVs with Samsung Brand are generally more expensive (£580 vs £450).

Display Technology: Most products in the result have LED Display Technology. Some have unspecified. TVs with LED Display Technology are generally less expensive (£370 vs £450).

Connectivity Technology: Most products in the result either have USB, WiFi, HDMI, or Ethernet Connectivity Technology. TVs with USB Connectivity Technology are generally less expensive (£355 vs £450).

HDR: Most products in the result have no HDR. A large proportion have HDR. TVs with no HDR are generally less expensive (£310 vs £450).

Figure 3: [alg2] Summary Example

The TV is the heart of your home cinema system. It's often the focal point of gatherings with friends or family, where you can catch up on the news, cheer for a sports team, enjoy the latest episode of that drama you love, or re-watch your favorite film. ... When you're looking for a new television at a great price with convenient delivery options, Amazon.co.uk is the place for you. You'll find a wide array of TVs that will fit both your needs and your budget.

Figure 4: An Excerpt from the Baseline Summary

to rank the summaries; "Please rank the summaries (#1 being most useful)" Then, they were asked 3 free text questions:

[Q1]: "From the summaries above, which part do you think is most useful? (please quote)"

[Q2]: "What would you like to see added to the summaries?"

[Q3]: "Which part do you think is not necessary?"

After the 2nd scenario was completed, we asked the participants to select 3 products they liked. Then we asked:

[Q4]: "When buying a TV which feature do you think is most important?"

[Q5]: "What information do you think should be in a summary?"

[Q6]: "What kind of summary would help you choose a good TV?"

For [Q4], participants could choose from a list with the following choices: price, screen size, supported content service, smart/internet, resolution, Freeview, connectivity (ports) and also could specify their own features.

Hypotheses: Our research hypotheses were:

[Hyp1]: Participants prefer the [alg2] summary over the baseline [amz] summary

[Hyp2]: Participants prefer the [alg2] summary over the [alg1] summary

4.2 Results

Summary Preference: The number of times each summary was ranked first, second or third in the 36 trials is as shown in Figure 5. The average ranking of each algorithm 1.47 [alg2], 1.81 [alg1], and 2.72 [amz] respectively.

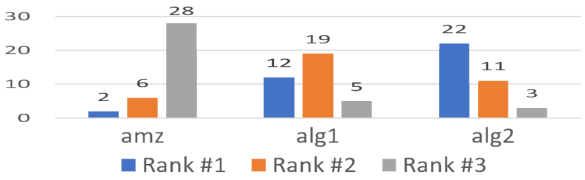

Figure 5: Ranking Counts of Each Algorithm

Out of all 36 trials, there were 31 (86.11%) where the participants preferred [alg2] over [amz] and 24 (66.67%) where the participants preferred [alg2] over [alg1].

A Friedman (1940) analysis of the rankings confirmed that the distributions of rankings were different for [amz], [alg1] and [alg2] (p-value of 2.8×10^{-7}). A post-hoc Friedman Aligned Ranks test (García et al., 2010) showed that [alg2] was significantly better than [amz] (p-value of 1.09×10^{-9}), thus confirming [Hyp1]. We could not confirm [Hyp2] as the p-value for [alg2] vs [alg1] was 0.104, though the numerical difference was in the expected direction.

Free Text Answer: Many responses (7 in total) asked for the summary to be short and precise or even bulleted. Furthermore, to [Q1] most participants found the price range and the relationship between price and features useful, which was supported by the data in the ranking. For [Q2], participants wanted to see product rating and other features, e.g. display frequency, model year or warranty added to the summary. They wanted to see some explanation of the technical terms and/or specification (e.g. what a smart TV is and what it can do). To [Q3], most participants did not find the Amazon summary useful and thought that it was not necessary. To [Q4], participants emphasized price (14 counts), screen size (11 counts), resolution (10 counts) and smart/internet feature (8 counts) when buying a TV. To the questions [Q5] and [Q6], participants thought that the features and their descriptions (including terminology explanations), how the features impact the price, user reviews, and information about warranty make a good summary.

5 Discussion and Future Work

Generalisation of our findings – which were based on only a very small set of scenarios – is tricky: we do not know whether they generalise to different kinds of products (e.g., groceries or paintings) and to product sets of different sizes (e.g. a set of just 3 products). However, our results suggest that customers find high-level product set summaries of the type we investigated more useful than Amazon's static product category overviews. This was further confirmed by the free text question where many participants quote substantial parts of [alg2] summary as being useful.

In future, we aim to experiment with refinements and extensions of [alg2]. For instance, in order to expand the algorithm work with various product domains, an automation of the analysis of hand-written reviews has to be implemented.

Additionally, based on participants' comments, technical information (as canned text) could also be included into the summary.

Since a number of readers pointed out that the summaries generated by [alg2] were too lengthy, the future version of the summary could be shorten (e.g., by omitting price comparisons in some cases). Some comments proposed that the summary should group together features that make the products different, separately from those the products have in common, this, as well, has a potential as a next feature to be experimented on.

In addition, to mimic more of human-written texts, approches to reduce the repetition in the generated text could be considered.

Finally, a more seamless integration of the summary to e-commerce websites could also be considered, maybe as a browser extension or a website wrapper.

Acknowledgment

The authors would like to thank the reviewers for their constructive comments and suggestions.

References

Milton Friedman. 1940. A comparison of alternative tests of significance for the problem of m rankings. *Ann. Math. Statist.*, 11(1):86–92.

Salvador García, Alberto Fernández, Julián Luengo, and Francisco Herrera. 2010. Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Inf. Sci.*, 180(10):2044–2064.

Stephanie Inglis, Ehud Reiter, and Somayajulu Sripada. 2017. Textually summarising incomplete data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 228–232, Santiago de Compostela, Spain. Association for Computational Linguistics.

Kittipitch Kuptavanich. 2018. Using textual summaries to describe a set of products. In *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys '18. ACM.

Prashant Mathur, Nicola Ueffing, and Gregor Leusch. 2017. Generating titles for millions of browse pages on an e-commerce site. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 158–167. Association for Computational Linguistics.

Kathleen R. McKeown. 1985. Discourse strategies for generating natural-language text. *Artificial Intelligence*, 27(1):1–41.

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

B. Shneiderman. 1996. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*, pages 336–343.

Kees Van Deemter. 2002. Generating Referring Expressions: Boolean Extensions of the Incremental Algorithm. *Computational Linguistics*, 28(1):37–52.