reviews associated to the 50 test queries. We denote the proposed approach and the old approach as "two-stage" and "one-stage", respectively. Due to the limited space, we only give a visual comparison of the two approaches on "image quality" in Figure 6. The upper figure shows the summarization of positive opinions and the lower figure shows that of negative opinions. From the figures we can see that the two-stage approach preserves fewer text segments as the result of filtering out many low-quality product reviews.

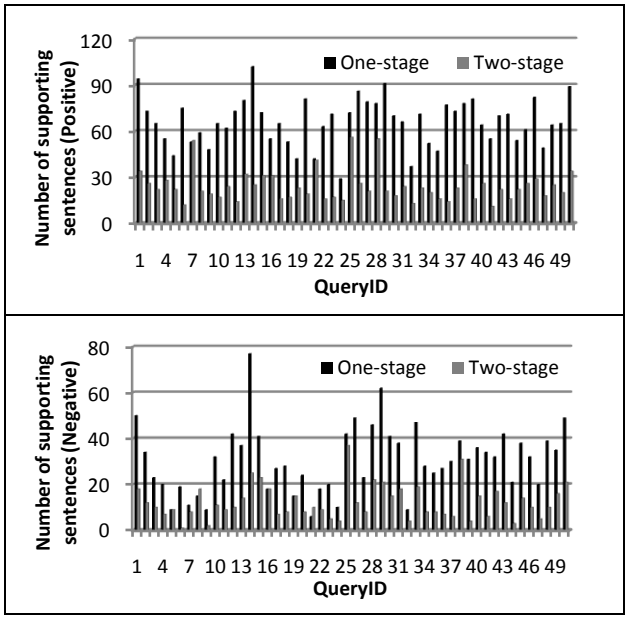

Figure 6. Summarization on "image quality"

To show the comparison on more features in a compressed space, we give the statistic ratio of change between two approaches instead. As for the evaluation measure, we define "*RatioOfChange*" (*ROC)* on a feature *f* as,

$$\mathrm{ROC}(f) = \frac{\mathrm{Rate}_{\mathrm{one-stage}}(f) - \mathrm{Rate}_{\mathrm{two-stage}}(f)}{\mathrm{Rate}_{\mathrm{one-stage}}(f)} \qquad (3)$$

where Rate $_*(f)$ is defined as,

$$\mathrm{Rate}_*(f) = \frac{|\mathrm{POS}(f)|}{|\mathrm{POS}(f)| + |\mathrm{NOS}(f)|} \qquad (4)$$

Table 4 shows some statistic results on *ROC* on five product features, namely "image quality"(IQ), "battery", "LCD screen" (LCD), "flash" and "movie mode" (MM). The values in the cells are the percentage of queries whose *ROC* is larger/smaller than the respective thresholds. We can see that a large portion of queries have big changes on the values of *ROC*. This means that the result achieved by the two-stage approach is substantially different from that achieved by the one-stage approach.

%Query	***RatioOfChange* (+)**					
	>0.30	**>0.25**	**>0.20**	**>0.15**	**>0.10**	**>0.05**
IQ	2%	4%	4%	10%	14%	22%
Battery	10%	14%	18%	30%	38%	50%
LCD	12%	18%	20%	22%	24%	28%
Flash	6%	10%	16%	20%	26%	42%
MM	6%	8%	8%	12%	18%	26%
%Query	***RatioOfChange* (-)**					
	<-0.30	**<-0.25**	**<-0.20**	**<-0.15**	**<-0.10**	**<-0.05**
IQ	4%	6%	10%	14%	18%	44%
Battery	2%	4%	4%	10%	14%	22%
LCD	4%	4%	8%	12%	22%	28%
Flash	4%	6%	8%	16%	18%	28%
MM	8%	10%	16%	18%	34%	42%

Table 4. *RatioOfChange* on five features

There is no standard way to evaluate the quality of opinion summarization as it is rather a subjective problem. In order to demonstrate the impact of the two-stage approach, we turn to external authoritative sources other than Amazon.com as the objective evaluation reference. We observe that CNET[2] provides a professional "*editor's review*" for many products, which gives a rating in the range of 1~10 on product features. 9 digital cameras out of the 50 test queries are found to have the editor's rating on "image quality" at CNET. We use this rating to compare with the results of our opinion summarization. We rescale the *Rate* scores obtained by both the one-stage approach and the two-stage approach into the range of 1-10 in order to perform the comparison.

Figure 7 provides the visual comparison. We can see that the result achieved by the two-stage approach has a much better (closer) resemblance to CNET rating than one-stage approach does. This indicates that our two-stage approach can achieve a more consistent summarization result to the professional evaluations by the editors. Although the CNET rating is not the absolute standard for product evaluation, it provides a professional yet objective evaluation of the products. Therefore, the experimental results demonstrate that our proposed approach could achieve more reliable opinion summarization which is closer to the generic evaluation from authoritative sources.

[2] http://www.cnet.com