Aspect-based summarization of pros and cons in unstructured product reviews

Florian Kunneman
Tilburg University
PO Box 90153
5000LE Tilburg, The Netherlands
f.a.kunneman@uvt.nl

Sander Wubben
Tilburg University
PO Box 90153
5000LE Tilburg, The Netherlands
s.wubben@uvt.nl

Emiel Krahmer
Tilburg University
PO Box 90153
5000LE Tilburg, The Netherlands
e.j.krahmer@uvt.nl

Antal van den Bosch
KNAW Meertens Institute
Oudezijds Achterburgwal 185
1024DK Amsterdam, The Netherlands
antal.van.den.bosch@meertens.knaw.nl

Abstract

We developed three systems for generating pros and cons summaries of product reviews. Automating this task eases the writing of product reviews, and offers readers quick access to the most important information. We compared SynPat, a system based on syntactic phrases selected on the basis of valence scores, against a neural-network-based system trained to map bag-of-words representations of reviews directly to pros and cons, and the same neural system trained on clusters of word-embedding encodings of similar pros and cons. We evaluated the systems in two ways: first on held-out reviews with gold-standard pros and cons, and second by asking human annotators to rate the systems' output on relevance and completeness. In the second evaluation, the gold-standard pros and cons were assessed along with the system output. We find that the human-generated summaries are not deemed as significantly more relevant or complete than the SynPat systems; the latter are scored higher than the human-generated summaries on a precision metric. The neural approaches yield a lower performance in the human assessment, and are outperformed by the baseline.

1 Introduction

Reviews posted on Web-based consumer platforms such as Amazon[1] and Trustpilot[2] form a valuable source of information prior to buying a product or hiring a service. Especially when a multitude of reviews of a specific product are gathered, the different experiences enable the consumer to obtain a well-informed conception of the qualities and deficiencies of the respective article. However, achieving a coherent view of a product tends to be challenging when a high volume of reviews diverge in focus and sentiment. A system that can summarize the reviews of a product would therefore be a valuable asset to these platforms. This work describes the implementation and evaluation of both a supervised and an unsupervised approach to aspect-based sentiment analysis, with a focus on Dutch product reviews of electric devices.

We create summaries that zoom in on the pros and cons of the product that are mentioned in a review. Platforms that include product reviews commonly employ a pros and cons template to enable the authors of reviews themselves to summarize their review text. Although such a procedure could lead to

[1] https://www.amazon.com/

[2] https://www.trustpilot.com/

Proceedings of the 27th International Conference on Computational Linguistics, pages 2219–2229
Santa Fe, New Mexico, USA, August 20-26, 2018.

appropriate summaries, only a minority of users make the effort to fill in the pros and cons.[3] On the one hand, this behavior underlines the need for a tool to generate a summary when the writer chooses not to. On the other hand, the review texts that do include pros and cons as explicated by the writer offers the train and test labels to employ a supervised learning framework. We will study the potential of using these user-generated summaries by training a neural-network implementation to generate the appropriate summary as output to the review text as input. This approach will be compared to a knowledge-driven approach based on syntactic patterns.

Aspect-based sentiment analysis from product reviews, of which the detection and summarization of pros and cons is an instance, is a challenging task due to the freedom that the user has when writing a review. In this case, the freedom applies to both the pros and cons the user chose to mention, and to the phrasings used to express these pros and cons. In addition, different types of products are reviewed with a focus on different aspects. For this reason, aspect-based sentiment analysis is commonly approached as a bounded task, with a focus on a specific product or service with predefined aspects. In contrast, we leverage the pros and cons that writers attach to their review, resulting in a more dynamic approach that can cope with any set of reviews, irrespective of the product or service, as long as example summaries are available. While the pros and cons as put forward by the users offers a gold standard for evaluation, we also include a human evaluation in which the qualities of both the system-generated summaries and the human-generated summaries are assessed.

This study offers the following contributions:

- We provide new insights into the utility of author-contributed pros and cons for automatically summarizing product reviews;
- We compare a neural learning approach with a linguistic pattern-driven approach;
- A novel evaluation is proposed that uses the author-assigned pros and cons as gold standard;
- The quality of the author gold standard is compared to system output through human evaluation.

After discussing related work, this paper will proceed with a description of the methods. Subsequently, the experimental set-up will be outlined, followed by the results. The paper ends with a conclusion and discussion.

2 Related Work

Studies in aspect-based sentiment analysis can roughly be divided into rule-based and machine learning-based approaches. Rule-based approaches implement searching for combinations of aspects and evaluations on the sentence level. Such combinations are commonly identified based on syntactic patterns (Poria et al., 2014; Zhang et al., 2014; Chinsha and Joseph, 2015; Rana and Cheah, 2015; Yan et al., 2015) while the evaluation is scored based on a sentiment lexicon (Chinsha and Joseph, 2015; Rana and Cheah, 2015; Poria et al., 2014), a PageRank-based procedure (Yan et al., 2015), or optimization based on review labels (Zhang et al., 2014). We apply a knowledge-driven approach to aspect-based sentiment analysis, comparable to the approach by Rana and Cheah (2015). Candidate phrases are identified based on predefined syntactic patterns and a sentiment lexicon is used to identify patterns that likely signify a pro or con.

Utilizing the pros and cons in product reviews as training data, we also apply a machine-learning based approach, using neural networks. Neural models have become popular for text classification tasks (Socher et al., 2013; Zhang et al., 2015; Conneau et al., 2016) in recent years. These models have the advantage that they generally require little feature engineering. The downside is that they are slow to train and test and require large amounts of training data. More recently, it has been shown that shallow neural models obtain performance on par with deep learning methods, while being much faster and less

[3] Of the reviews that are used in our study, only 30% include pros and/or cons. A full overview of the reviews is given in Table 2

data hungry. The higher representational power of these deep methods is not required for relatively simple tasks such as sentiment classification (Joulin et al., 2016).

Few studies on aspect-based sentiment analysis make use of the pros and cons that are given by the writers themselves to summarize their review. Closest to our work in this respect is the contribution by Kim and Hovy (2006), who align pros and cons with the most likely related sentence in their review text by a maximum entropy model. The output is used as training data to predict the pros and cons for reviews that were not summarized as such by the writer. In comparison to Kim and Hovy (2006), we leverage author-based pros and cons both as training labels and as test labels during evaluation. In addition, we evaluate these pros and cons along with system-generated pros and cons, thereby providing insight into the quality of these kinds of labels for training and evaluating aspect-based sentiment analysis of product reviews.

3 Method

3.1 System 1: Syntactic patterns of subjective phrases (SynPat)

The knowledge-driven SynPat system[4] tracks and matches syntactic cues towards aspect evaluations in reviews (Rana and Cheah, 2015). The system consists of two stages. First, a review is scanned for particular syntactic patterns based on a set of rules. Second, each candidate that is found during the first stage is matched with a lexicon of subjective words to confirm that the phrase indicates valence.

Part-of-speech tags and grammatical phrases are extracted from each review text by means of Frog[5], an off-the-shelf NLP system for Dutch (Van den Bosch et al., 2007). We matched the automatic analyses with a set of syntactic patterns of which the lexical realizations (the words associated with the PoS-tags) may reflect statements on the valence of an aspect. Examples of patterns and example lexical realizations are given in Table 1.

To assess the presence and direction of the valence in the lexical realizations (henceforth referred to as phrases) found with matching syntactic patterns, we matched them to the Duoman subjectivity lexicon (Jijkoun and Hofmann, 2009). This a lexicon of nouns and adjectives rated by human annotators as 'Very negative', 'Negative', 'Neutral', 'Positive' and 'Very positive'. We labeled each phrase with a positive or very positive word as 'pro', and each phrase with a negative or very negative word as 'con'. Phrases in which no word has a positive or negative valence in Duoman lexicon were discarded, as well as phrases that matched an equal amount of positive and negative words. A phrase that carries a combination of a 'positive' and 'very negative' is labeled as 'con', while phrases with a 'very positive' and 'negative' word are labeled as 'pro'.

Syntactic pattern	Example (Dutch)	Example English	Valence word	Pro/con
ADJ	fantastisch	fantastic	fantastic	Pro
ADJ-N	klein zuigmondje	small nozzle	small	Con
ADJ-N-N	heerlijk kopje espresso	delicious cup of espresso	delicious	Pro
ADJ-ADJ-N	onhandige extra handeling	inconvenient additional action	inconvenient	Con
ADJ-V	mooi uitgevoerd	nicely rendered	nicely	Pro
ADJ-Prep-N	goed van smaak	good flavor	good	Pro

Table 1: Examples of syntactic patterns extracted in the SynPat system (ADJ=Adjective, N=Noun, V=Verb, Prep=Preposition), their lexical realization (phrase)

3.2 System 2: Shallow neural classification

To explore the possibility of training a classification system end-to-end without feature engineering we choose to train a shallow neural network on the aspect summarization task. We treat reviews as bags of words and feed word embeddings to the neural classifier. Given the constructed sentence vector, the

[4]The implementation of the system can be found on github: `https://github.com/fkunneman/Product_review_summary/blob/master/synpat.py`

[5]`http://applejack.science.ru.nl/languagemachines/software/frog/`

classifier tries to predict the correct labels. Here, the labels are the human generated aspect sentences. We treat each aspect as a potential class to predict. The chosen architecture is similar to Joulin et al. (2016) which is a modification of the cbow model by Mikolov et al. (2013), but instead of predicting a hidden word, a hidden class is predicted. We use softmax to compute the probability distribution over the aspect classes. The model is trained using stochastic gradient descent and a linearly decaying learning rate. To pretrain the model we use 320-dimensional word embeddings derived from the corpora from the web (COW) introduced in Tulkens et al. (2016).

3.3 System 3: Shallow neural classification with clustering

Because users can state positive and negative aspects in any way they want, we end up with a large number of classes in a long-tail distribution. From inspection of the development data we observe that many of the different classes are in fact paraphrases of the same semantic content ('great coffee', 'coffee tastes great', etc.) To try to alleviate this problem, we add a preprocessing step where we cluster the aspect summaries based on averaging over word embedding vectors. Using k-means clustering we then cluster the summaries into distinct clusters and take the centroid as the label. We then try to predict these labels using the same shallow neural classifier.

4 Experimental Set-up

4.1 Data

We crawled product reviews from the Dutch platform 'kieskeurig.nl', which features consumer reviews for a diversity of electric devices. A screen shot of a typical consumer review is presented in Figure 1. Typically, such a review consists of a written text, a summary of the written text in the form of pros and cons, a review score and, occasionally, scores on predefined aspects such as design and ease of use. We are interested in the review text and the accompanying pros and cons.

Figure 1: Example of a review placed on kieskeurig.nl. 'Pluspunten' are pros, 'Minpunten' cons.

We crawled the reviews for the four most frequently reviewed product types: smartphones, vacuum cleaners, deep fryers and espresso machines.[6] The statistics of the resulting dataset are presented in Table 2. We observe that a significant part of the reviews are written without summing up explicit pros and cons. In addition, pros are more common than cons. About 35% of the reviews are accompanied by one or more explicated pros, while cons are present for 22% of the reviews.

4.2 Procedure

The review texts summarized with pros and cons were utilized as instances to train our systems with, and to test them intrinsically on held-out data from the same source. In order to tune the neural networks, we

[6]Due to privacy regulations, we are not allowed to publicly share this data.

Device	# reviews	# reviews with at least one pro	# reviews with at least one con	# reviews with both pros and cons	# reviews with either a pro or a con
Smartphone	5,133	1,007	703	697	1,012
Deep fryer	4,040	1,526	1,039	1,019	1,536
Vacuum cleaner	2,623	1,026	730	715	1,041
Espresso machine	2,282	979	682	666	986
Total	14,078	4,538	3,154	3,097	4,575

Table 2: Overview of the review data set

Device	# train	# dev	# test	# total
Smartphone	809	101	102	1,012
Deep fryer	1,228	154	154	1,536
Vacuum cleaner	832	104	105	1,041
Espresso machine	793	99	100	986

Table 3: Overview of the data used in the experimental set-up

split the reviews into training (80%), development (10%) and test (10%) instances. The resulting dataset is presented in Table 3.

The unsupervised SynPat system was applied directly to the test set. The hyperparameters of the two shallow neural systems were set using the development set. We trained for 30 epochs using a learning rate of 0.5 and the softmax loss function. For k-means clustering we set $k = 100$.

We compared the performance of the three systems to the performance of a baseline system based on string matching. This works as follows. All pros and cons contained in the training reviews of a specific device are saved as possible descriptions of unseen reviews. For a given unseen review, all known pros and cons from the training reviews are matched with the text. The baseline selects as summary the pros and cons of which an above-threshold proportion of words is seen in the text. The threshold value is empirically decided based on the development set: the threshold that leads to the pros and cons that are closest to the actual given pros and cons is selected and used for the predictions on the test set.[7]

4.3 Evaluation

We evaluated the systems in two ways: a gold standard assessment based on the pros and cons given by the writers of the reviews and a human evaluation in which the pros and cons given by the systems as well as the writers themselves are assessed by human annotators.

4.3.1 Evaluation 1: Gold standard

For the gold standard evaluation, we assessed the pros and cons extracted by a system as their closeness in characters to the author-based pros and cons.[8] Due to the variance in these pros and cons, which are given by the author in a free-text input field, it is infeasible to make direct matches between two sets of pros and cons. Slight grammatical differences between target and prediction would be assessed as a false prediction in such a procedure. Alternatively, we made use of the FuzzyWuzzy string matching library for python.[9] At the basis of this implementation is a computation of the edit distance between two strings using the Ratcliff–Obershelp pattern recognition algorithm (Ratcliff and Metzener, 1988). In this algorithm, the edit distance between two strings is computed by dividing the number of matching characters with the total number of characters. The FuzzyWuzzy library is particularly adequate for

[7]The implementation of the baseline can be found on github: `https://github.com/fkunneman/Product_review_summary/blob/master/baseline.py`

[8]The implementation of evaluation 1 can be found on github: `https://github.com/fkunneman/Product_review_summary/blob/master/evaluation1.py`

[9]`https://github.com/seatgeek/fuzzywuzzy`

assessing the distance between longer phrases, as it expands the core pattern recognition algorithm with two heuristics. The first is to overrule the strong penalization of two strings with a differing length by assigning a more positive score if one of the two strings is a substring of the other. The second heuristic assures that the order of word tokens is not of influence to the edit distance, by separately assessing the intersection of tokens between two phrases as two unordered sets.

Using the FuzzyWuzzy string matching procedure, the quality of the systems are assessed in the following way, for each review in the test set:

1. The pros and cons suggested by the system and the ones given by the user are aligned such that a predicted pro is matched with only one gold standard pro, the most similar one, and a predicted con is matched with only one gold standard con. Two predicted pros or cons can not be aligned with the same gold standard pro or con; the pair with the highest similarity is aligned.

2. Apart from string matching, we employed a heuristic for matching values of pros and cons that refer to 'Nothing'. This is the most common value (especially to the cons field), and is often given by not filling in anything or using words like 'geen' ('none') and 'nog niets gevonden' ('nothing found yet'). The heuristic assures that such values were matched. Such matches were given the optimal similarity score.

3. Aligned pairs with a similarity that surpasses a given threshold are scored as 'true positives'. Other pairs are scored as false positives (for system-generated pros and cons) and false negatives (for user-generated pros and cons). Pros and cons that are not aligned are also scored as 'false positives' and 'false negatives'

4. A precision, recall and F1-score for the summary of each single review is calculated based on the counts of true positives, false positives and false negatives. After all reviews in the test set are processed this way, an aggregate precision, recall and F1-score of the system is calculated.

In this evaluation procedure, true positives are assigned when the similarity between a target pro or con and a predicted pro or con surpasses a given threshold. To obtain insight into the influence of the threshold value, we will run the evaluation with different thresholds that cover the total spectrum of possible edit distance scores. The range of thresholds spans from 0 (no similarity at all) to 100 (complete similarity). Precision, recall and F1-score is calculated separately per review rather than adding up all true positives, false negatives and false positives, so as to avoid the influence of imbalance in the number of pros and cons per review.

4.3.2 Evaluation 2: Human assessment

In addition to a gold standard evaluation, we asked human annotators to assess the quality of review summaries. Apart from the pros and cons generated by the baseline system, SynPat, Neural and Neural_Clust, we included the pros and cons given by the authors themselves (the gold standard pros and cons) in this evaluation. This can give insight into the absolute quality of the gold standard, and its quality with respect to the systems. This kind of experimentally collected (human) judgments is often used in the evaluation of automatically generated text (Gatt and Krahmer, 2018).

We randomly selected twenty reviews from the test set, equally distributed over the four devices 'vacuum cleaner', 'deep fryer', 'espresso machine' and 'smartphone' (e.g.: five reviews per device). Each review was included five times in the evaluation set: linked with the pros and cons generated by each of the four systems as well as the pros and cons given by the author of the review. This led to a total of 100 instances. We divided them into five chunks of twenty instances to be presented to human annotators, which would include all twenty unique review texts presented with the pros and cons of one of the five sources (equally distributed over the twenty reviews). Each chunk of twenty review-system combinations was assessed by the same ten human annotators, enabling us to calculate the agreement between each annotator pair.

Annotators were recruited through the Radboud Research Participation System[10], a university-based platform that helps researchers finding student participants for their studies. For the assessment of twenty review texts, participants could choose to receive one student credit or a voucher worth ten Euros. In total, fifty human annotators assessed the quality of the pros and cons of twenty review texts. For each of these instances, the annotator was asked to assess both the relevance and the completeness of the given pros and cons with respect to the review text. Relevance was assessed by deciding for each of the given pros and cons whether they are actually discussed (literally or paraphrased) in the review text. Completeness was assessed by indicating, on a scale from 1 to 7, the extent to which the given summary covers all pros and cons that are mentioned in the review text. Relevance can be seen as an estimator of precision; completeness is a human-assessed approximation of recall.

A completeness and relevance score per review-system combination was computed as the average of the ten human assessments. The score per system is calculated as the average of these averages.

5 Results

An overview of the number of (aligned) predictions by system is presented in Table 4. Recall that the pros and cons in the predicted and gold standard summaries are aligned based on string matching and a heuristic for the most common phrases to express that there is no pro or con. In the latter case, the highest similarity score of 100 is assigned to the alignment. We therefore present the average similarity of aligned pros and cons with and without the ones aligned based on this 'empty' heuristic.

The baseline is characterized by a high number of predictions and alignments. The proportion of predictions that are aligned is, however, the lowest. Most of these alignments are based on string matching, which has the highest average similarity (50.5) of all systems. The SynPat system makes the least predictions, but yields a relatively high proportion of alignments and a high number of reviews with an aligned prediction. The average similarity of the alignments is slightly lower than the baseline. The Neural system yields the lowest number of aligned pros and cons based on string matching (468), but the highest number of alignments based on the 'empty' heuristic. The percentage of reviews with an alignment is low at 81%, which is due to the high amount of falsely predicted 'empties'. This seems to be the default option for Neural when certainty for other pros or cons is low. The Neural_clust system makes aligned predictions for most of the reviews, at an average string similarity of 44.43.

System	Predicted pros and cons	Aligned pros and cons (string matching)	Aligned pros and cons ('empty' heuristic)	# reviews aligned (% of total)	Avg. similarity	Avg. similarity (without 'empty')
Baseline	3,877	1,469	79	432 (94%)	52.50	50.05
SynPat	1,712	1,101	83	442 (96%)	51.41	48.42
Neural	1,844	468	159	375 (81%)	58.58	44.99
Neural_clust	1,844	1,228	109	456 (99%)	48.96	44.43

Table 4: Overview of predictions and alignments by system for all reviews (# target reviews: 461, # target pros and cons: 2060, similarity scaled 0-100).

5.1 Gold standard

A similarity threshold decides which pros and cons predicted by a system are assessed as matches (true positives) with a gold standard pro or con. The resulting F1-scores by threshold value are presented in Figure 2 (vacuum cleaner), Figure 3 (smartphone), Figure 4 (deep fryer) and Figure 5 (espresso machine).

The curves for most systems are showing a steep drop in performance around a threshold of 30, after which the scores are more stable between 50 and 70 and steadily decrease towards an F1-score of about 0.1 at 100. The SynPat and Neural_clust systems yield the highest performance between a threshold value of 0 and 30–40 for all systems, indicating that these systems make the most predictions that are

[10] https://radboud.sona-systems.com

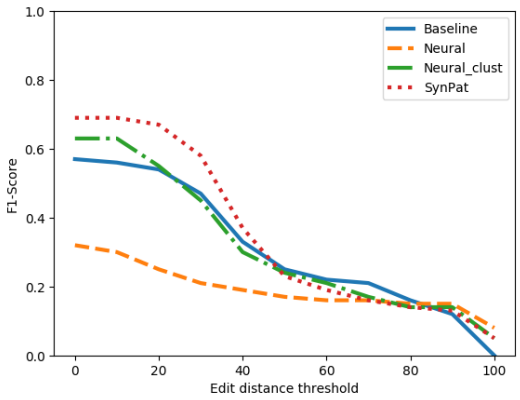

Figure 2: Performance by similarity threshold on vacuum cleaner reviews

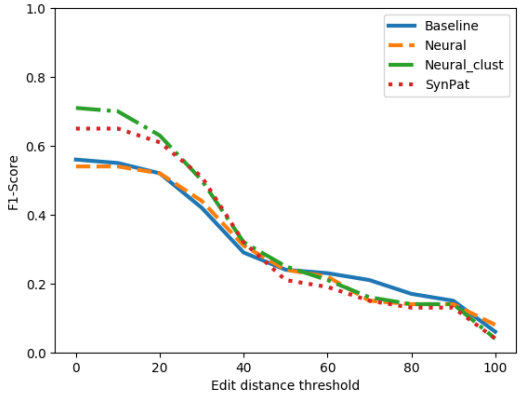

Figure 3: Performance by similarity threshold on smartphone reviews

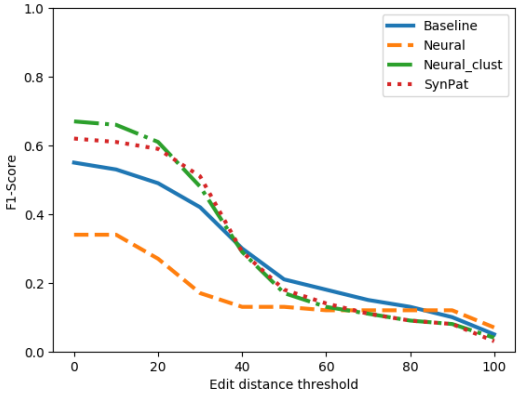

Figure 4: Performance by similarity threshold on deep fryer reviews

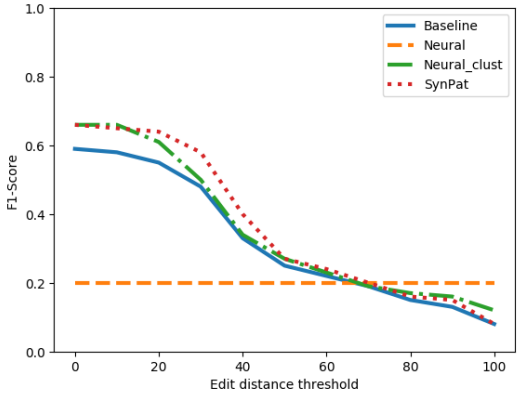

Figure 5: Performance by similarity threshold on espresso machine reviews

somewhat similar to one of the gold standard pros or cons. The F1-scores of all systems are closer to each other at distance thresholds of over 40. The baseline system outperforms the other systems between the threshold values of 60 and 80. The Neural system is generally outperformed at threshold values below 60 (except for the smartphone reviews). The clustering of pros and cons seems to be useful for this approach. Upon closer inspection, the flat line for the Neural system on the espresso machine reviews (Figure 5) is due to the exclusive prediction of 'Nothing' for cons.

5.2 Human assessment

A report of the human assessments of the pros and cons given by each system, as well as the writers of the reviews themselves, is given in Table 5. A Cohen's Kappa (Cohen, 1960) agreement was calculated between the relevance assessments of pros and cons of each pair of annotators. The average agreement on relevance was $\kappa = 0.56$. The agreement over the completeness assessment of pros and cons between each annotator pair was calculated as a weighted Cohen's Kappa. The average agreement for completeness was $\kappa = 0.30$. These Kappa-values suggest a moderate agreement between human judges.

Inspection of Table 5 reveals that the reviewers' summaries were judged to be the most complete at an average score of 4.60, although the differences with the average score of SynPat (4.06) and the Baseline (3.90) appear to be small. Apparently, reviewers themselves often do not cover all pros and cons that they mention in their review text. The relevance of their pros and cons (0.61) is lower than the best system, SynPat (0.67), although this difference is again small. The baseline system lags behind in terms of relevance. The scores of the two Neural systems are lower than the scores of the other systems for both completeness and relevance, and are close to each other.

We conducted two one-way Analyses of Variance (ANOVAs) to test for statistical significance of the

	Completeness	Relevance
Baseline	3.90 (0.90)	0.44 (0.17)
SynPat	4.06 (1.15)	0.67 (0.25)
Neural	2.74 (0.86)	0.25 (0.14)
Neural_clust	2.37 (0.62)	0.18 (0.10)
Reviewers summary	4.60 (1.13)	0.61 (0.25)

Table 5: Outcomes of the human assessment of the pros and cons generated by each system and the writers of the reviews themselves for 20 reviews. Completeness is rated on a scale from 1 to 7, and is reported as the average completeness of the 20 reviews (standard deviation between brackets). Relevance is rated as the proportion of pros and cons that are deemed relevant to a review text, on a scale from 0 to 1, reported as the average relevance of the 20 reviews (standard deviation between brackets).

completeness and relevance assessments. The ANOVA for completeness confirmed that the systems were not rated equally on this dimension, $F(4, 95) = 23.98, p < .001$. All pairwise comparisons (using the Tukey HSD method) were statistically significant at $p < .05$, except the comparisons between the human gold standard and SynPat, between human gold standard and baseline, and between SynPat and baseline (confirming that the three highest rated approaches performed statistically equally well) and between the two neural systems (confirming that neither neural method outperformed the other in terms of completeness). The ANOVA for relevance similarly showed that not all systems performed equally well, $F(4, 95) = 18.61, p < .001$, with all pairwise comparisons (Tukey HSD) statistically signifcant at $p < .05$, except the ones between the human gold standard and SynPat (confirming that SynPat performs on a par with human reviewers), and between the two neural methods.

5.3 Analysis

We set out to investigate the strength of each system by counting the overlap of their predictions for each of the test reviews. To focus on the most sensible predictions, we only selected the predictions with a similarity of > 50 with one of the gold standard pros and cons. An overview of the overlap is presented in a Venn diagram in Figure 6.

The diagram shows a predominantly disjoint pattern, with only 3.6% of the correctly predicted pros and cons generated by all four systems. The combination of three systems with most correct predictions that a fourth system has not generated (5.8%) is formed by the baseline, SynPat and Neural_clust. In most of these cases, the Neural system has falsely predicted 'empty' as con. The most overlapping combination of systems is the combination baseline–SynPat, with 120 predictions (11.6%). Both systems are based on matching phrases, while the Neural systems operate at the sentence level. Apparently, a good part of the pros and cons can best be detected on the phrase level.

6 Conclusion and Discussion

We presented three systems for generating pros and cons summaries of product reviews: a system based on syntactic phrases and a valence lexicon (SynPat), a neural-network-based system trained on bag-of-words in the review text as input and the pros and cons listed by the writer as output (Neural), and the same system trained on clusters of similar pros and cons based on word embeddings (Neural_clust). These systems were applied on reviews of vacuum cleaners, smartphones, deep fryers and espresso machines. We find that the pros and cons assigned by the authors themselves are not deemed by human annotators as more relevant or complete than the summaries extracted by the SynPat system. In fact, the SynPat system is assessed as offering more relevant pros and cons than the human summaries (though not significantly more). The neural-network approaches yield a lower performance in the human assessment, and are also outperformed by the baseline.

The lower performance of the neural-network approaches is likely due to the formulation of the task as a classification task of a text (either encoded as a bag-of-words or as centroids of word embedding clusters) into many different output classes. Moving from this discriminatory approach to a generative

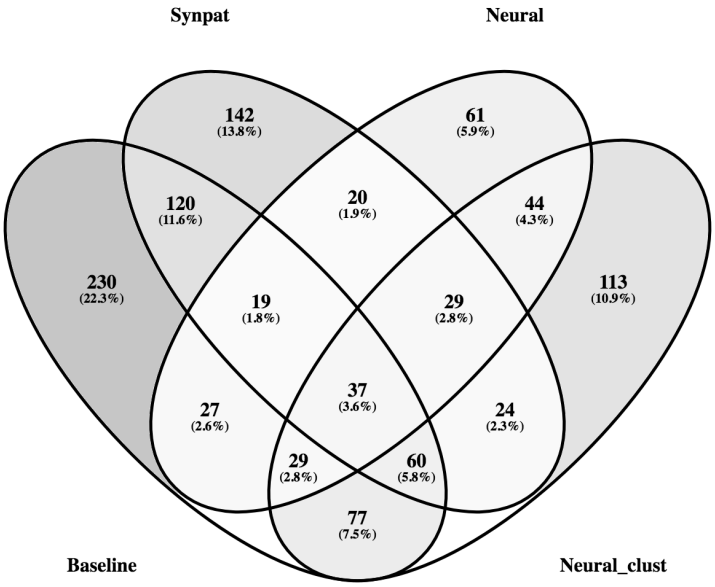

Figure 6: Venn diagram of the overlap between predicted pros and cons by each system which match the gold standard pros and cons at a similarity score of > 50.

approach may be a next step to attain better performances.

Our results show that human performance is imperfect as well. The SynPat system is assessed as not significantly worse than humans, and is even slightly more precise. The upper bound that the best systems appear to touch in Figures 2 to 5, an F1-score of about 0.7, must therefore be bounded by the variance and imperfect completeness (recall) and relevance (precision) of the human summaries we took as gold standard. In line with this finding, the moderate Cohen's Kappa scores that describe the agreement of human annotators on assessing the link between pros and cons and a review text ($\kappa = 0.56$ for relevance and $\kappa = 0.30$ for completeness) underlines the subjective nature of this task. In future work, subjectivity can be reduced (and thereby human agreement increased) by employing a stricter definition of when elements in a review text reflect a pro or con.

Finally, in future work we intend to move beyond extractive summarization of pattern-matched phrases. A strong assumption of the SynPat system is that pros and cons are consecutive word sequences that can be extracted literally from the full review, while pros and cons could be expressed in ways that transcend mere word n-grams. This calls for a richer encoding of the full review text, likely between the level of n-grams and the full bag-of-words. The sentence may be a good domain to target.

Acknowledgements

This work is part of the research programmes Dis- cussion Thread Summarization for Mobile Devices and The Automated Newsroom, which are financed by the Netherlands Organisation for Scientific Research (NWO). We thank the reviewers for their valuable comments.

References

TC Chinsha and Shibily Joseph. 2015. A syntactic approach for aspect based opinion mining. In *Semantic Computing (ICSC), 2015 IEEE International Conference on*, pages 24–31. IEEE.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Alexis Conneau, Holger Schwenk, Loïc Barrault, and Yann Lecun. 2016. Very deep convolutional networks for natural language processing. *arXiv preprint arXiv:1606.01781*.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Valentin Jijkoun and Katja Hofmann. 2009. Generating a non-english subjectivity lexicon: Relations that matter. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 398–405. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Soo-Min Kim and Eduard Hovy. 2006. Automatic identification of pro and con reasons in online reviews. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 483–490. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Soujanya Poria, Erik Cambria, Lun-Wei Ku, Chen Gui, and Alexander Gelbukh. 2014. A rule-based approach to aspect extraction from product reviews. In *Proceedings of the second workshop on natural language processing for social media (SocialNLP)*, pages 28–37.

Toqir Ahmad Rana and Yu-N Cheah. 2015. Hybrid rule-based approach for aspect extraction and categorization from customer reviews. In *9th International Conference on IT in Asia (CITA)*. IEEE.

John W Ratcliff and David E Metzener. 1988. Pattern-matching-the gestalt approach. *Dr Dobbs Journal*, 13(7):46.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Stephan Tulkens, Chris Emmery, and Walter Daelemans. 2016. Evaluating unsupervised dutch word embeddings as a linguistic resource. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Antal Van den Bosch, Bertjan Busser, Sander Canisius, and Walter Daelemans. 2007. An efficient memory-based morpho-syntactic tagger and parser for Dutch. In P. Dirix, I. Schuurman, V. Vandeghinste, and F. Van Eynde, editors, *Computational Linguistics in the Netherlands: Selected Papers from the Seventeenth CLIN Meeting*, pages 99–114, Leuven, Belgium.

Zhijun Yan, Meiming Xing, Dongsong Zhang, and Baizhang Ma. 2015. Exprs: An extended pagerank method for product feature extraction from online consumer reviews. *Information & Management*, 52(7):850–858.

Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 83–92. ACM.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657.