Maejo Int. J. Sci. Technol. **2016**, *10*(02), 142-153

Maejo International
Journal of Science and Technology

ISSN 1905-7873
Available online at www.mijst.mju.ac.th

Full Paper

Exploring linguistic structure for aspect-based sentiment analysis

Nuttapong Sanglerdsinlapachai [1, *], Anon Plangprasopchok [2] and Ekawit Nantajeewarawat [1]

[1] School of Information, Computer and Communication Technology, Sirindhorn International Institute of Technology, Thammasat University, Pathumthani, 12000, Thailand
[2] Innovation and Engineering Research Unit, National Electronics and Computer Technology Centre, Pathumthani, 12120, Thailand
* Corresponding author, e-mail: nuttapongs@gmail.com

Received: 13 October 2015 / Accepted: 19 May 2016 / Published: 25 May 2016

Abstract: Aspect-based sentiment analysis is a task that extracts relevant sentiments of a specific aspect. An opinion text is usually composed of views on different aspects of an entity. By investigating the sources of errors, we observe that a scoring method at the level of elementary discourse units (EDUs) highly contributes to the accuracy of sentiment classification at the aspect level. Score aggregation can be improved by considering linguistic structures between EDUs in a hierarchical manner. We propose a new score aggregation strategy that incrementally aggregates sentiment scores from EDUs to local segments and from local segments to an aspect. The experimental results on a product review dataset demonstrate that our new score aggregation method improves the performance of sentiment classification at the aspect level. At the EDU level, calculation of polarity scores using an all-term average yields better performance compared to score calculation based on opinion phrases extracted by using term dependencies.

Keywords: aspect-based sentiment analysis, linguistic structure, opinion phrase, dependency pattern, rhetorical structure theory

INTRODUCTION

Sentiment analysis is a method of extracting sentiments from natural language texts [1, 2]. The output of the analysis varies and depends on a user's definition such as polarity (negative, neutral or positive) and subjectivity (objective or subjective) [2]. Sentiment analysis occurs at different levels of the text structure, namely words, phrases, clauses, sentences, or the entire document [2, 3]. Sentiment analysis at the *aspect* level [2, 4, 5] is challenging because a text usually contains different opinions on many aspects. For example, a user may post a review of the mobile phone, mentioning the pros and cons of its several features (aspects), viz. signal, weight, price, etc.

An aspect-based sentiment analysis performs two tasks [2]. The first is aspect extraction, which identifies the aspects of a given text. An aspect is identified or found via descriptive statistics, e.g. term frequencies [6], and topic modelling techniques, e.g. Latent Dirichlet Allocation (LDA) [7]. Jo and Oh [4] applied LDA at the sentence level in order to detect aspects and its respective sentiment words. They proposed a unified model that combines aspect extraction and sentiment analysis. Moghaddam and Ester [5] also proposed a framework that employs LDA. In particular, an opinion phrase, a pair of terms and their modifiers, was first extracted by using manually defined rules. The phrase then became a substitute source for term occurrences (bag-of-words) in the LDA process. After processing LDA with bag-of-phrases, top opinion phrases for a document were obtained and aspects, along with associated sentiments, were then extracted from those opinion phrases.

The second task is the sentiment analysis of aspects obtained from the first task. Two major types of techniques, machine-learning-based and lexicon-based, may perform this task. In the machine-learning-based approach [8, 9], the text content is segmented into a bag of words; then a machine-learning technique, e.g. Support Vector Machine (SVM) [8], is applied to learning term occurrence patterns for each polarity class. On the other hand, the lexicon-based method uses a list of sentiment-carrying words or lexicon corpus, e.g. SentiWordNet [10], as a main resource. It seeks to achieve a proper linear combination of the term scores that represents an overall sentiment. Chamlertwat et al. [11] proposed a simple yet efficient method that sums polarity scores of all terms to represent the document's sentiment and applied the method to a Twitter-posted dataset. Negation terms (e.g. not, no) are used as triggers to flip the sentiment polarity of a particular post. However, there is a common drawback to these two approaches. Since texts are disintegrated to a bag-of-word representation, linguistic structures of the content are neglected. This means that performance declines when we apply these methods to texts with rich linguistic structures.

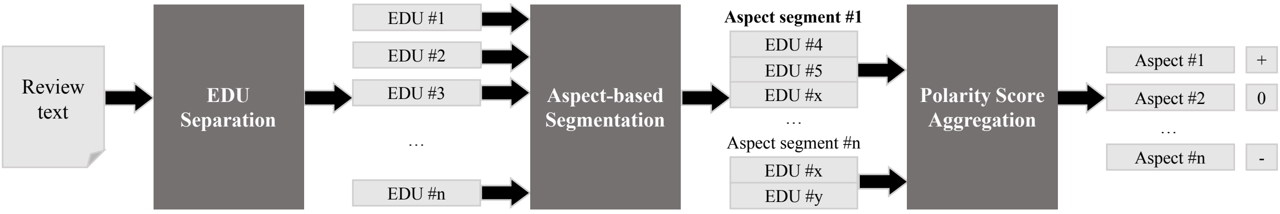
Some researchers [12, 13] were interested in the relationships between text clauses. Specifically, they hypothesised that the relationships could boost the sentiment classification accuracy by appropriately aggregating sentiments across interrelated clauses. Rhetorical Structure Theory (RST) [14], which defines several types of clause relations, has been adopted in many studies including ours [15]. In the sentence-level sentiment analysis by Chenlo et al. [16], RST was applied to disintegrating a sentence into nucleus and satellite sections. After calculating polarity scores separately, scores from the two parts were merged and the merging was weighted according to the relationship of the nucleus and satellite sections. In our previous work on aspect-based sentiment analysis [15], we used relations in RST to determine aspect-dependent elementary discourse units (EDUs). The score representing the polarity of the aspect was then calculated by averaging the polarity scores of the collected EDUs. Our work performed reasonably well at the average f-measure of 0.765.

Recently, some researchers proposed hybrid methods, combining the machine-learning-oriented method and lexicon-based method with linguistic structural information. Chenlo and Losada [17] extended their previous lexicon-based approach [16] by treating extracted relations in RST as a set of features for SVM and Logistic Regression to classify sentiments. Wachsmuth et al. [18] proposed a new feature called sentiment flow pattern for recording polarity changes of a set of EDUs linked by RST relations. The approach predicts the overall polarity of a given text according to recognised patterns and is operable in texts with multiple aspects.

In this paper we extend our previous work [15] by incorporating two kinds of linguistic structures: an RST tree structure relating EDUs and a term-dependency structure within an EDU. The hierarchical structure of an RST tree is used for aggregating polarity scores of EDUs that are relevant to a given aspect. The score averaging scheme performs well for polarity aggregation. The dependency structure of terms within an EDU is exploited for choosing important polarity terms representing the aspect polarity of the EDU. Comparisons with methods for calculating polarity scores without using linguistic structures are made.

PRELIMINARY STUDY

In this section we briefly explain our previous work [15], in which our aspect-based sentiment analysis was applied to a dataset consisting of 108 mobile phone reviews, collected from CNET [19]. Each review is concerned with one or more aspects in a predetermined collection of 13 manually defined aspects, viz. screen, application, network, system, camera, capacity, power/battery, sensor, accessory, size, hardware/body, sound, and price. As shown in Scheme 1, an RST parser [20] is first employed to segment a textual review into EDUs and connect them using RST relations. Aspects relevant to the review are identified by matching their predetermined keywords with terms in the EDUs. An EDU containing at least one keyword of an aspect is called a *key EDU* for the aspect. The key EDUs for an aspect *asp* and their adjacent EDUs with respect to RST relations are grouped into an *aspect segment* for *asp*. The polarity of an aspect is then determined by averaging the polarity scores of the EDUs in the associated aspect segment. The hierarchical structure of EDUs has not been employed for polarity score calculation.



Scheme 1. Overview of our previous work [15]

Our previous method [15] was shown to significantly improve the naive polarity score calculation method presented by Chamlertwat et al. [11], in which all words occurring in an entire review text are used for finding the polarity of an aspect. However, there is still room for improvement considering the resulting f-measure values. In this paper we investigate two types of possible improvement: (1) score aggregation at the aspect level and (2) polarity score calculation at the EDU level.

ASPECT-LEVEL SCORE AGGREGATION

Our product review dataset contains 330 aspect segments with a total of 1,537 EDUs. To investigate score aggregation at the aspect level, three human annotators were requested to manually label each individual EDU in the dataset by selecting polarity scores from the choices -1, -0.5, 0, 0.5 and 1, which denote ‘strongly negative’, ‘weakly negative’, ‘neutral’, ‘weakly positive’ and ‘strongly positive’ respectively. The average annotated score is taken as its actual label. The polarity of an aspect segment is then calculated by averaging the scores of all EDUs in the segment.

An accuracy of 82.4% is obtained. This occurs even when the manually annotated EDU-level scores are used in the calculation; 17.6% of all aspects are still incorrectly classified. To improve the classification accuracy, we propose to divide an aspect segment into smaller segments, which we call *local aspect segments*. We calculate their polarity scores independently. The overall polarity of the aspect segment is then determined by averaging the polarity scores of all local segments contained in it.

Local Aspect Segments

An aspect may be expressed in many text segments that are not necessarily adjacent to each other. Each individual segment has an independent polarity about the aspect. For example, consider a portion of a review in Scheme 2. The segment 'Switching from another operating system' is classified as neutral for the 'system' aspect, while the segment 'I love not only the camera, but the Windows Phone operating system' holds a positive polarity for the same aspect. A proper aggregation is required to produce the overall polarity of the aspect.

> "The complete mobile package, great hardware, apps, more". Switching from another operating system, I wasn't sure what to expect (other than a good camera). I love not only the camera, but the Windows Phone operating system. Wish it were less expensive, but with this kind of hardware, I'm willing to pay for the extras.

Scheme 2. Sample review showing non-adjacent EDUs concerning 'system' aspect

We define a *local aspect segment* as either (1) a span of RST elements that has a key EDU as its nucleus or (2) a key EDU that appears as a satellite of an RST relation. Diagram 1 illustrates local aspect segments in a product review, where squares labelled with 0-9 represent EDUs and bold-border squares (i.e. squares with the labels 0, 1, 3, 4 and 5) represent key EDUs. The diagram contains six local aspect segments relevant to four aspects; details are listed in Table 1. The aspect segment for the 'camera' aspect, as well as that for the 'system' aspect, consists of two local aspect segments. Local aspect segments for different aspects may involve the same group of EDUs, e.g. local aspect segments *Hardware*-1 and *Application*-1 in Table 1.

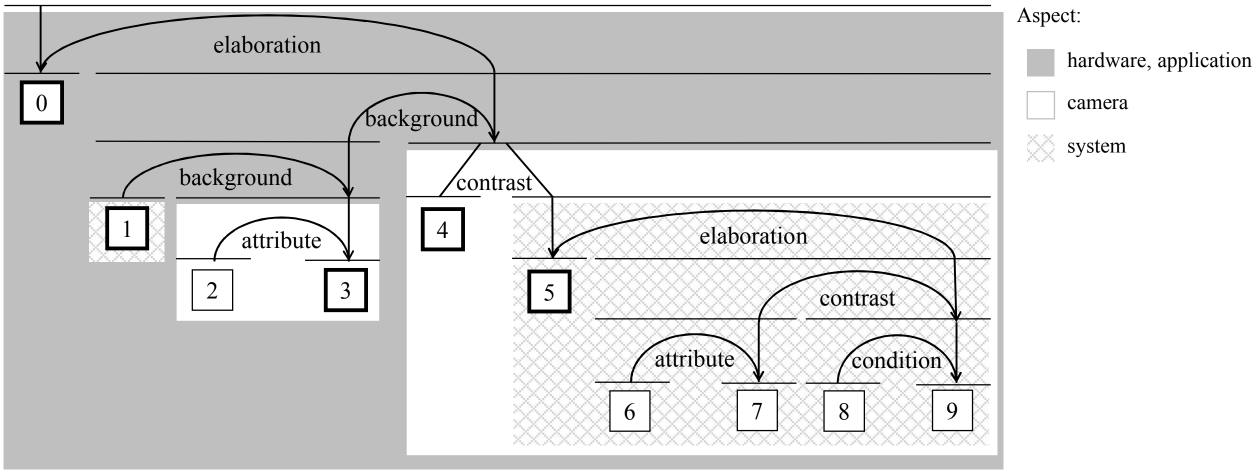


Diagram 1. RST relation tree depicting aspect segments and their components

Table 1. Local aspect segments extracted from sample review in Diagram 1

Aspect	Local aspect segment	EDU
'hardware'	*Hardware*-1	0, 1, 2, 3, 4, 5, 6, 7, 8, 9
'application'	*Application*-1	0, 1, 2, 3, 4, 5, 6, 7, 8, 9
'camera'	*Camera*-1	4, 5, 6, 7, 8, 9
	Camera-2	2, 3
'system'	*System*-1	5, 6, 7, 8, 9
	System-2	1

Our dataset contains 583 local aspect segments. For evaluation purposes, all local aspect segments were manually annotated with their actual polarity labels. By averaging the labels of local aspect segments relevant to each aspect, the polarity of aspect segments can be classified with an improved accuracy of 85.8%. These preliminary results show promise in developing an approach for automatically calculating polarity scores at the level of local aspect segments. The resulting scores can then be used for calculating the overall polarity at the aspect level.

Score Aggregation for Local Aspect Segments

To aggregate the polarity scores of the EDUs in a local aspect segment, a basic approach is to simply take their average value and ignore their linguistic structure. This method is referred to as the *All-EDU averaging scheme*. Another method called *Top-RST-node averaging scheme* considers the top-level RST relation in a local aspect segment. When a local aspect segment contains multiple EDUs, its polarity score is obtained by averaging the score of the key EDU (i.e. the nucleus of the top-level relation) and the average score of the remaining EDUs in the local aspect segment. Referring to Diagram 1 and Table 1, for example, the polarity score of the local aspect segment *System*-1 is obtained by averaging the score of EDU 5 and the value obtained by averaging the scores of EDUs 6, 7, 8 and 9.

We evaluated the two methods using the manually labelled scores of EDUs as input data and the manually labelled scores of local aspect segments as ground-truth data. The All-EDU averaging scheme classifies 90.2% of the segments correctly, while the Top-RST-node averaging scheme gives a higher accuracy of 92.8%.

EDU-LEVEL SCORE CALCULATION

For automatic calculation of EDU-level polarity scores, a basic method called *All-term averaging method* was used in our previous work [15]. This method uses SentiWordNet [10] to determine the polarity scores of all individual terms occurring in an EDU and takes their average score as the polarity score of the EDU. The resulting score is flipped to the opposite sign (i.e. negative or positive) if at least one negation term appears in the EDU. We investigate an alternative EDU-level score calculation method, i.e. score calculation based on opinion phrases extracted using the internal linguistic structure of an EDU.

Score Calculation Based on Opinion Phrases

Terms occurring in the same EDU may not be relevant to the same aspect. For example, consider the EDU 'If you are looking for great hardware without a ton of unused apps on it that you cannot remove'. This EDU contains two aspects, i.e. 'hardware' and 'application'. The polarity score of the 'hardware' aspect might be considered from the term 'great', which is its only modifier, while other terms such as ‘without’, ‘unused’ and ‘cannot remove’ are relevant to the ‘application’ aspect. Taking such a multiple-aspect issue into consideration, Zhan and Li [21] introduced the notion of an *opinion phrase*, which is a pair of a clue term of an aspect (*heading*) and a modifier term (*modifier*), and applied it to the sentiment classification model. In order to apply opinion phrases to EDU-level score calculation, we extract opinion phrases whose headings are predetermined keywords of a given aspect, and then calculate the polarity score of the aspect by averaging the polarity scores of all modifiers in the extracted opinion phrases. This EDU-level scoring method is called the *Opinion-phrase-based method*.

Opinion Phrases Extraction

To extract opinion phrases from a given text, Moghaddam and Ester [5] proposed nine rules, referred to as *ME-rules* (Scheme 3). These rules are derived from dependencies or grammatical relations between terms according to *Stanford Typed Dependencies Manual* [22], i.e. adjectival modifier (*amod*), nominal subject (*nsubj*), adjectival complement (*acomp*), copular complement (*cop*), direct object (*dobj*), conjunction (*conj_and*), negation (*neg*) and noun compound modifier (*nn*). Some rules are constrained by parts of speech, e.g. a noun (*N*), a verb (*V*) or an adjective (*A*), while others consider the similarity between the heading (*h*) or the modifier (*m*) of a previously extracted opinion phrase and a term in a grammatical relation. For example, consider the EDU ‘The battery life of this device is amazing’, which contains the relations *nn*(*life*, *battery*), *nsubj*(*amazing*, *life*) and *cop*(*amazing*, *is*). When the ME-3 rule is applied to this EDU, the opinion phrase ⟨*life*, *amazing*⟩ is extracted. When the ME-8 rule is applied to this extracted opinion phrase and the relation *nn*(*life*, *battery*), the opinion phrase ⟨*battery_life*, *amazing*⟩ is further extracted.

ME-1: $amod(N, A) \rightarrow \langle N, A\rangle$
ME-2: $nsubj(V, N) + acomp(V, A) \rightarrow \langle N, A\rangle$
ME-3: $nsubj(A, N) + cop(A, V) \rightarrow \langle N, A\rangle$
ME-4: $nsubj(V, N') + dobj(V, N) \rightarrow \langle N, V\rangle$
ME-5: $\langle h_1, m\rangle + conj\_and(h_1, h_2) \rightarrow \langle h_2, m\rangle$
ME-6: $\langle h, m_1\rangle + conj\_and(m_1, m_2) \rightarrow \langle h, m_2\rangle$
ME-7: $\langle h, m\rangle + neg(m, x) \rightarrow \langle h, -m\rangle$
ME-8: $\langle h, m\rangle + nn(h, N) \rightarrow \langle N\_h, m\rangle$
ME-9: $\langle h, m\rangle + nn(N, h) \rightarrow \langle h\_N, m\rangle$

Scheme 3. ME-rules for extracting opinion phrases using term-dependency relations

EXPERIMENTS AND RESULTS

We compared polarity scoring methods by conducting experiments on three levels: the EDU level, the local aspect segment level and the aspect level.

Sentiment Classification at EDU Level

At the EDU level, the All-term averaging method was compared with the Opinion-phrase-based method on two sets of EDUs, called the *Key-EDU set* and the *All-relevant-EDU set*. The first set contains all key EDUs (totalling 673 key EDUs) for the 13 aspects considered in our mobile phone review dataset. The second set contains all EDUs (totalling 1,537 EDUs) appearing in the 330 aspect segments in the dataset. For comparison, the annotated polarity scores and the computed polarity score of each EDU were aligned to one of the three polarity orientations, i.e. negative, neutral or positive, depending on whether they are less than, equal to or greater than zero. The results of this experiment are shown in Table 2.

Table 2. Comparison between All-term averaging and Opinion-phrase-based methods

EDU set	**Actual polarity**	**Number of EDUs**	**Number of correctly classified EDUs**	
			All-term averaging	**Opinion-phrase-based**
Key-EDU set	Negative	155	92 (59.4%)	32 (20.6%)
	Neutral	144	19 (13.2%)	111 (77.1%)
	Positive	374	322 (86.1%)	144 (38.5%)
	Overall	673	433 (64.3%)	286 (42.5%)
All-relevant-EDU set	Negative	312	179 (57.4%)	32 (10.3%)
	Neutral	545	106 (19.4%)	512 (93.9%)
	Positive	680	560 (82.4%)	144 (21.2%)
	Overall	1,537	845 (55.0%)	688 (44.8%)

The All-term averaging method outperforms the Opinion-phrase-based method at the overall accuracy of 64.3% and 55.0% on the Key-EDU set and the All-relevant-EDU set respectively. The actual positive EDUs are more correctly classified than the actual negative EDUs by both methods on both EDU sets. A possible reason is that most positive EDUs contain only positive sentiment terms without negative terms, whereas negative EDUs contain either pure negative sentiment terms or positive terms with negation words, making the sentiment classification more difficult. Considering the actual neutral EDUs, the Opinion-phrase-based method obviously outperforms the All-term averaging method. Most neutral EDUs that are misclassified by the All-term averaging method contain mixtures of positive and negative sentiment terms with their calculated polarity score being not exactly equal to zero. The accuracy of the Opinion-phrase-based method depends heavily on the coverage of the ME-rules. For example, a non-neutral EDU is classified as neutral when no ME-rule is applicable to it or no aspect keyword appears as the heading of an extracted opinion phrase. The rule coverage issue will be discussed further in the next section.

Score Aggregation for Local Aspect Segments

We next evaluated the methods for aggregating the polarity scores of EDUs at the level of local aspect segments, i.e. the All-EDU averaging scheme and the Top-RST-node averaging scheme. Table 3 shows the evaluation results. Using manually annotated EDU-level scores, both schemes yield the accuracy values of at least 90% with the accuracy of Top-RST-node averaging scheme being higher. Using automatic EDU calculation, both schemes yield the same performance. The score aggregation using the EDU-level scores obtained from the All-term averaging method outperforms that using the EDU-level scores obtained from the Opinion-phrase-based method.

Score Aggregation at Aspect Level

To evaluate the usage of local aspect segments in sentiment classification of aspect segments, two aspect-level score aggregation approaches were considered: (1) score aggregation without local aspect segments and (2) score aggregation with local aspect segments. For the first approach, two methods for combining EDU-level scores were considered: (1a) averaging the scores of key EDUs only and (1b) averaging the scores of all relevant EDUs in an aspect segment (which was the method used in our previous work [15]). For the second approach, the two score aggregation schemes for local aspect segments described earlier were applied, i.e. (2a) the All-EDU averaging scheme and (2b) the Top-RST-node averaging scheme. The polarity score of an aspect segment was then determined by averaging the resulting scores of all local aspect segments contained in it. The experimental results are shown in Table 4.

Table 3. Comparison of methods of score aggregation for local aspect segments

Score aggregation method	**Number of correctly classified local aspect segments**		
	Using manually annotated EDU-level scores	**Using automatic EDU-level score calculation**	
		All-term averaging	**Opinion-phrase-based**
All-EDU averaging scheme	526 (90.2%)	369 (63.3%)	229 (39.3%)
Top-RST-node averaging scheme	541 (92.8%)	369 (63.3%)	229 (39.3%)

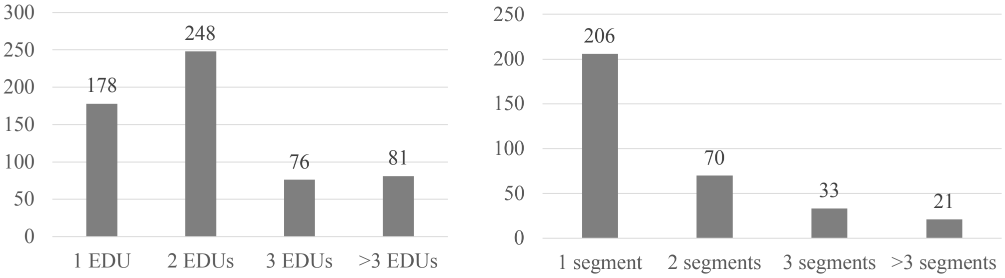
Table 4. Comparison of methods of score aggregation for aspect segments

Score aggregation method		**Number of correctly classified aspect segments**		
		Using manually annotated EDU-level scores	**Using automatic EDU-level score calculation**	
			All-term averaging	**Opinion-phrase-based**
Without local aspect segments	(1a) Averaging scores of key EDUs only	263 (79.7%)	233 (70.6%)	114 (43.6%)
	(1b) Averaging scores of all relevant EDUs	272 (82.4%)	239 (72.4%)	114 (43.6%)
With local aspect segments	(2a) All-EDU averaging scheme	275 (83.3%)	239 (72.4%)	146 (44.2%)
	(2b) Top-RST-node averaging scheme	275 (83.3%)	239 (72.4%)	146 (44.2%)

The results show that the score aggregation with local aspect segments is more accurate than that without local aspect segments. For score aggregation without local aspect segments, the results support the claim in our previous work [15] in that the use of key EDUs in combination with their relevant EDUs improves the accuracy of sentiment classification at the aspect level, compared to the use of key EDUs alone. For score aggregation with local aspect segments, the All-EDU averaging scheme and the Top-RST-node averaging scheme yield the same accuracy. Considering the EDU-level score calculation, the All-term averaging method achieves higher accuracy than the Opinion-phrase-based method.

Further Examination

To provide an insight into the results in Table 3 and 4, we examined our product review dataset by considering the number of EDUs in each local aspect segment and the number of local aspect segments in each aspect segment, and investigated the cases when the local aspect segments or aspect segments were incorrectly classified. Figure 1(a) shows that most local aspect segments (approximately 73%) consist of one EDU or two EDUs. When applied to such local aspect segments, the All-EDU averaging scheme and the Top-RST-node averaging scheme always yield the same results. When the application of these two averaging schemes to the local aspect segments that contain more than two EDUs were examined, we find that: (1) when the Opinion-phrase-based method is used at the EDU level, the two schemes give the same classification results, and (2) when the All-term averaging method is used at the EDU level, the two schemes classify 12 local aspect segments differently, with a half of them being misclassified by each scheme. As a result, using automatic EDU-level score calculation, the All-EDU averaging scheme and the Top-RST-node averaging scheme give the same accuracy in our dataset (cf. Table 3).



(a) Number of EDUs per local aspect segment (b) Number of local aspect segments per aspect

Figure 1. Distribution of EDUs in local aspect segments (a) and distribution of local aspect segments in aspect level (b)

For the aspect level, we specifically examined the application of score aggregation without local aspect segments using all relevant EDUs and that with local aspect segments using the All-EDU averaging scheme (i.e. (1b) and (2a) respectively in Table 4). Figure 1(b) shows that more than 60% of aspect segments consist of only one local aspect segment. When applied to the aspect segments of this type, the two aspect-level score aggregation methods always give the same results. When applied to the remaining aspect segments using the EDU-level All-term averaging method, the two aspect-level score aggregation methods classify four aspect segments differently, with two aspect segments being misclassified by each method. The two aspect-level methods therefore yield the same accuracy in our dataset (cf. the column 'All-term averaging' in the second and third rows of Table 4).

EXTENSIONS OF OPINION-PHRASE-BASED METHOD

From the experimental results at the EDU level (cf. Table 2), the Opinion-phrase-based method yields lower overall accuracy compared to the All-term averaging method. We investigated two possible extensions of the Opinion-phrase-based method, i.e. (1) improving the coverage of extraction rules by addition of new rules and (2) applying the Opinion-phrase-based method in combination with the All-term averaging method.

Extension of Extraction Rule

The ME-rules in Scheme 3 are applicable to only 47.6% of the key EDUs in our product review dataset. By adding new extraction rules, useful opinion phrases can be additionally extracted and employed for polarity score calculation. For example, consider the key EDU 'With normal use the battery works fine', which contains the grammatical relations *amod*(*use*, *normal*), *pobj*(*with*, *use*), *det*(*works*, *the*), *nn*(*works*, *battery*), *npadvmod*(*fine*, *works*) and *amod*(*use*, *fine*). By adding a rule for the *npadvmod* relation (noun phrases as adverbial modifier), the opinion phrase ⟨*works*, *fine*⟩ can be extracted. By applying the ME-8 rule in Scheme 3 to this opinion phrase and the relation *nn*(*works*, *battery*), the opinion phrase ⟨*battery_works*, *fine*⟩ indicating positive polarity for the 'power/battery' aspect can be derived.

To discover more useful rules, we first applied the *sequential covering technique* [23] to our dataset to identify significant dependency relations for extraction of opinion phrases whose headings contain aspect keywords. From nine additionally obtained relations, 18 additional extraction rules were then constructed. Using the extended set of extraction rules, opinion phrases can be extracted from 83.3% of the key EDUs in our dataset. As a result, the overall accuracy of the Opinion-phrase-based method is increased to 61.2% on the Key-EDU set and 52.6% on the All-relevant-EDU set (compared with the overall accuracy of 42.5% and 44.8% respectively in Table 2).

Combining Opinion-phrase-based and All-term Averaging Methods

A non-neutral EDU is misclassified as neutral by the Opinion-phrase-based method when no opinion phrase is extracted from it. To reduce the possibility of such misclassification, the All-term averaging method is used in combination with the Opinion-phrase-based method. Figure 2 shows the distributions of the EDU-level polarity scores calculated using the All-term averaging method, indicating that the obtained scores of actual neutral EDUs are likely to be close to zero. A threshold can thus be set for predicting a neutral EDU based on the All-term averaging method when no opinion phrase is extracted. Let E be a given EDU. Let $score_{all}(E)$ denote the score of E calculated by the All-term averaging method and $score_{opi}(E)$ denote that calculated by the Opinion-phrase-based method. Using a predetermined threshold Th, the score of E, denoted by $score(E)$, is calculated by combining $score_{opi}(E)$ and $score_{all}(E)$ as follows:

- $score(E) = score_{opi}(E)$ if $score_{opi}(E) \neq 0$.
- $score(E) = score_{all}(E)$ if $score_{opi}(E) = 0$ and $|score_{all}(E)| > Th$.
- $score(E) = 0$ if $score_{opi}(E) = 0$ and $|score_{all}(E)| \leq Th$.

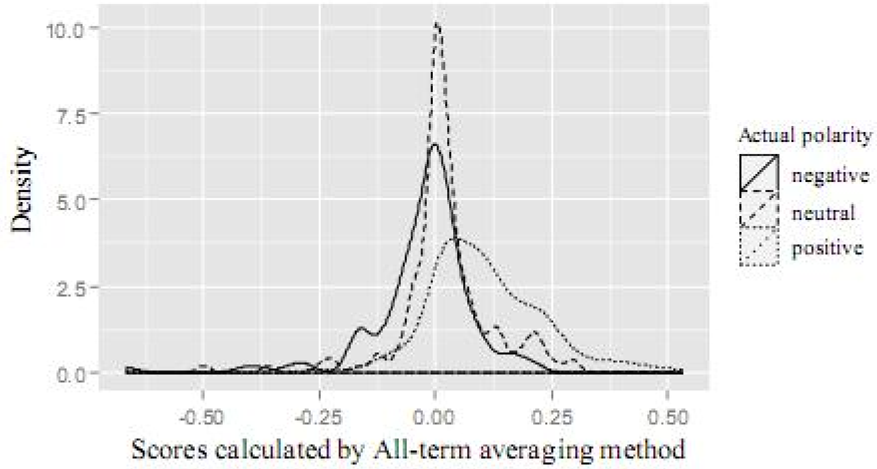


Figure 2. Distributions of EDU-level polarity scores calculated by All-term averaging method

With the threshold value of zero, this method improves the overall accuracy from 42.5% to 63.8% on the Key-EDU set and from 44.8% to 54.8% on the All-relevant-EDU set (cf. Table 2) when the original ME-rules in Scheme 3 are used. Using the extended set of extraction rules with the same threshold value improves the overall accuracy to 64.4% on the Key-EDU set and 55.0% on the All-relevant-EDU set. An appropriate threshold value in the range 0.0-0.5 is selected by applying the *grid search* method. Using the original ME-rules in Scheme 3 with the threshold value of 0.03, this method yields the overall accuracy of 60.3% on the Key-EDU set and 56.9% on the All-relevant-EDU set. Using the extended set of extraction rules with the threshold value of 0.05, the method yields the overall accuracy of 65.5% on the Key-EDU set and 61.6% on the All-relevant-EDU set.

CONCLUSIONS

Compared with the aspect-level polarity score aggregation by directly averaging the manually annotated EDU-level scores, score aggregation by averaging the annotated scores of intermediate-level local aspect segments has improved the overall polarity classification accuracy from 82.4% to 85.8% in our product review dataset. Considering automatic EDU-level scoring, the aspect-level aggregation with local aspect segments using the All-term averaging method has yielded a better accuracy by 28.2% compared to using the Opinion-phrase-based method. At the EDU level, with an extended set of extraction rules and an appropriate threshold value, a combination of the All-term averaging method and the Opinion-phrase-based method has yielded a better performance compared to the use of each individual method alone.

ACKNOWLEDGEMENTS

This work was partially supported by (1) the National Research University Project of Thailand, Office of the Higher Education Commission and (2) the Centre of Excellence in Intelligent Informatics, Speech and Language Technology and Service Innovation (CILS), Thammasat University.

REFERENCES

1. B. Pang and L. Lee, “Opinion mining and sentiment analysis”, *Found. Trends Inform. Retriev.*, **2008**, *2*, 1-135.
2. B. Liu and Z. Lei, “A survey of opinion mining and sentiment analysis”, in “Mining Text Data” (Ed. C. C. Aggarwal and C. X. Zhai), Springer, Boston (MA), **2012**, Ch.13.
3. W. Medhat, A. Hassan and H. Korashy, “Sentiment analysis algorithms and applications: A survey”, *Ain Shams Eng. J.*, **2014**, *5*, 1093-1113.
4. Y. Jo and A. H. Oh, “Aspect and sentiment unification model for online review analysis”, Proceedings of 4th ACM International Conference on Web Search and Data Mining, **2011**, Kowloon, Hong Kong, pp.815-824.
5. S. Moghaddam and M. Ester, “On the design of LDA models for aspect-based opinion mining”, Proceedings of 21st ACM International Conference on Information and Knowledge Management, **2012**, Maui, USA, pp.803-812.
6. M. Hu and B. Liu, “Mining and summarizing customer reviews”, Proceedings of 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, **2004**, Seattle, USA, pp.168-177.
7. D. M. Blei, A. Y. Ng and M. I. Jordan, “Latent dirichlet allocation”, *J. Mach. Learn. Res.*, **2003**, *3*, 993-1022.
8. R. Moraes, J. F. Valiati and W. P. G. Neto, “Document-level sentiment classification: An empirical comparison between SVM and ANN”, *Expert Syst. Appl.*, **2013**, *40*, 621-633.

9. S. Rustamov, E. Mustafayev and M. A. Clements, "Sentiment analysis using neuro-fuzzy and hidden Markov models of text", Proceedings of IEEE Southeastcon 2013, **2013**, Jacksonville, USA, pp.1-6.
10. A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining", Proceedings of 5th International Conference on Language Resources and Evaluation, **2006**, Genoa, Italy, pp.417-422.
11. W. Chamlertwat, P. Bhattarakosol, T. Rungkasiri and C. Haruechaiyasak, "Discovering consumer insight from twitter via sentiment analysis", *J. Univers. Comput. Sci.*, **2012**, *18*, 973-992.
12. C. Zirn, M. Niepert, H. Stuckenschmidt and M. Strube, "Fine-grained sentiment analysis with structural features", Proceedings of 5th International Joint Conference on Natural Language Processing, **2011**, Chiang Mai, Thailand, pp.336-344.
13. L. Polanyi and M. van den Berg, "Discourse structure and sentiment", Proceedings of 11th International Conference on Data Mining Workshops, **2011**, Vancouver, Canada, pp.97-102.
14. W. C. Mann and S. A. Thompson, "Rhetorical structure theory: Toward a functional theory of text organization", *Text Interdiscip. J. Study Discourse*, **1988**, *8*, 243-281.
15. N. Sanglerdsinlapachai, A. Plangprasopchok and E. Nantajeewarawat, "Exploiting rhetorical structures to improve feature-based sentiment analysis", Proceedings of 18th International Computer Science and Engineering Conference, **2014**, Khon Kaen, Thailand, pp.180-185.
16. J. M. Chenlo, A. Hogenboom and D. E. Losada, "Rhetorical structure theory for polarity estimation: An experimental study", *Data Knowl. Eng.*, **2014**, *94*, 135-147.
17. J. M. Chenlo and D. E. Losada, "An empirical study of sentence features for subjectivity and polarity classification", *Inform. Sci.*, **2014**, *280*, 275-288.
18. H. Wachsmuth, M. Trenkmann, B. Stein and G. Engels, "Modeling review argumentation for robust sentiment analysis", Proceedings of 25th International Conference on Computational Linguistics, **2014**, Dublin, Ireland, pp.553-564.
19. CNET, "Phone reviews", http://www.cnet.com/topics/phones/products (Accessed: Mar. 28, **2016**).
20. D. A. duVerle and H. Prendinger, "A novel discourse parser based on support vector machine classification", Proceedings of the Joint Conference of 47th Annual Meeting of the Association for Computational Linguistics and 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing: Vol. 2-Vol. 2, **2009**, Suntec, Singapore, pp.665-673.
21. T.-J. Zhan and C.-H. Li, "Semantic dependent word pairs generative model for fine-grained product feature mining", in "Advances in Knowledge Discovery and Data Mining" (Ed. J. Z. Huang, L. Cao and J. Srivastava), Springer-Verlag, Berlin-Heidelberg, **2011**, Ch.38.
22. M.-C. de Marneffe, B. MacCartney and C. D. Manning, "Generating typed dependency parses from phrase structure parses", Proceedings of 5th International Conference on Language Resources and Evaluation, **2006**, Genoa, Italy, pp.449-454.
23. T. M. Mitchell, "Machine Learning", McGraw-Hill, New York, **1997**, pp.275-282.