Knowledge-Based Systems 187 (2020) 104825

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Deep learning models and datasets for aspect term sentiment classification: Implementing holistic recurrent attention on target-dependent memories[☆]

Hyun-jung Park[a], Minchae Song[b], Kyung-Shik Shin[c,*]

[a] *Management Research Center, Ewha Womans University, Seoul, Republic of Korea*
[b] *Major in Big Data Analytics, Ewha Womans University, Seoul, Republic of Korea*
[c] *School of Business, Ewha Womans University, Seoul, Republic of Korea*

ARTICLE INFO

Article history:
Received 18 January 2019
Received in revised form 12 June 2019
Accepted 27 June 2019
Available online 3 July 2019

Keywords:
Aspect-based sentiment analysis
Sentiment classification
Deep learning
LSTM
GRU
Attention

ABSTRACT

An essential challenge in aspect term sentiment classification using deep learning is modeling a tailor-made sentence representation towards given aspect terms to enhance the classification performance. To seek a solution to this, we have two main research questions: (1) Which factors are vital for a sentiment classifier? (2) How will these factors interact with dataset characteristics? Regarding the first question, harmonious combination of location attention and content attention may be crucial to alleviate semantic mismatch problem between aspect terms and opinion words. However, location attention does not reflect the fact that critical opinion words usually come left or right of corresponding aspect terms, as implied in the target-dependent method although not well elucidated before. Besides, content attention needs to be sophisticated to combine multiple attention outcomes nonlinearly and consider the entire context to address complicated sentences. We merge all these significant factors for the first time, and design two models differing a little in the implementation of a few factors. Concerning the second question, we suggest a new multifaceted view on the dataset beyond the current tendency to be somewhat indifferent to the dataset in pursuit of a universal best performer. We then observe the interaction between factors of model architecture and dimensions of dataset characteristics. Experimental results show that our models achieve state-of-the-art or comparable performances and that there exist some useful relationships such as superior performance of bi-directional LSTM over one-directional LSTM for sentences containing multiple aspects and vice versa for sentences containing only one aspect.

1. Introduction

In the era of a hyper-connected society as today, people are increasingly expressing or retrieving opinions about a variety of products, organizations, and political issues via online channels such as blogs, forums, and e-commerce sites. Such vast amount of data accumulated on the Web has enormous practical value to enhance the quality of decision-making of organizations and individuals [1–5]. However, manual analysis of such a sheer volume of data will be almost infeasible. Thus, sentiment analysis dealing with how to automatically derive sentiment polarities toward various objects is attracting growing attention from both academic and industrial communities [4,6–10]. Sentiment analysis can be classified into three categories depending on the level of analysis: document, sentence, and aspect. Document and sentence level analyses determine the overall sentiment polarity of the whole document and sentence, respectively. Aspect level analysis, referred to as Aspect-Based Sentiment Analysis (ABSA) in this work, provides more fine-grained sentiment polarities toward each aspect of an entity described in a sentence [11–18]. For example, a review sentence about a laptop entity "CPU is satisfactory, but battery life is the worst". can be processed by ABSA to evaluate 'CPU' as positive and 'battery life' as a definite negative. Classifying the overall polarity, let us say, by a sentence level analysis, as negative makes us overlook critical opinions about CPU and battery life, resulting in a loss of valuable business opportunities.

Meanwhile, deep learning has demonstrated state-of-the-art performances in a variety of Natural Language Processing (NLP) tasks such as machine translation [19,20], question answering

[☆] No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.knosys.2019.06.033.

[*] Corresponding author.
E-mail addresses: sparrowpark@ewha.ac.kr (H. Park), mc.song@ewha.ac.kr (M. Song), ksshin@ewha.ac.kr (K. Shin).

https://doi.org/10.1016/j.knosys.2019.06.033
0950-7051/© 2019 Elsevier B.V. All rights reserved.

[21,22], and text summarization [23]. It has also been applied to the ABSA area. Deep learning for ABSA has the advantage of alleviating the burden of laborious feature engineering or building sentiment lexicon involved in the traditional machine learning techniques [24–26], although it is still in its infancy. Studies on ABSA utilizing deep learning have been conducted in four main tasks of ABSA: aspect category detection [27–30], aspect sentiment classification [27,29,31], aspect term extraction [9,32,33], and aspect term sentiment classification [31,32,34–38]. Aspect category detection, as a multi-class and multi-label classification problem, addresses identifying the aspect categories a sentence describes among predefined aspect categories. There can be multiple aspect labels for the same sentence. This task may be further classified depending on whether entities are known [39] or unknown [29], or whether aspects of entities are implicit or explicit [27,40]. Aspect sentiment classification as a multi-class classification problem infers sentiment polarity toward the aspect category extracted through the preceding aspect category detection. Aspect term extraction as a sequence labeling problem deals with identifying an arbitrary number of aspect terms contained in a sentence [32]. In the above sentence, 'CPU' and 'battery life' can be aspect terms corresponding to the aspect or aspect category of "CPU" and "battery", respectively. An aspect or aspect category can be represented by one or more than one aspect terms. Unless described otherwise, we assume, for simplicity, there exist a few consecutive aspect terms corresponding to an aspect or aspect category. Lastly, aspect term classification as a multi-class classification problem aims to determine sentiment polarity of a sentence concerning known aspect terms. These tasks have been explored somewhat independently due to their challenging nature except that a few studies have addressed more than one task in sequence or parallel [29,32,41].

This work focuses on the aspect term sentiment classification task for which many exciting models are being published thanks to the presence of aspect terms. To guarantee satisfactory classification performance, modeling a tailor-made sentence representation according to the concerned aspect terms is an essential problem of this task. To seek a solution to this, we have the following two main research questions. First, which factors are vital for an ABSA classifier to derive an informative sentence representation? Second, how will these factors interact with dataset characteristics toward classification performance?

Regarding the first question, inducing a sentiment classifier to pay more attention to the right words for the given aspect terms will be indispensable. For this purpose, the location attention mechanism which considers the distance between the concerned aspect terms and each context word can be a straightforward way [34,36,42]. With only location attention, however, relevant opinion words far from the aspect terms can be ignored. This limitation can be mitigated by the content attention mechanism that reflects semantic relatedness between the given aspect terms and each context word [31,34,36,43,44]. These two types of attention mechanisms, i.e. location attention and content attention, can act complementarily to reduce semantic mismatch problem between aspect terms and opinion words.

One thing to note about location attention is that it treats both sides of the concerned aspect terms equal without reflecting the fact that opinion words relevant to the aspect terms usually come before or after the aspect terms in a sentence. For instance, "CPU is satisfactory, but battery life is the worst". can be divided into "CPU is satisfactory, but battery life" and "battery life is the worst". when the focused aspect terms are 'battery life'. The crucial opinion words for 'battery life' are included in the right segment while the left segment does not look directly relevant to the aspect terms. Thus, it seems reasonable to form two different networks for left and right segments and then connect intermediate outputs from each part to compose an ultimate sentence representation.

By the way, one thing about content attention is that it needs to be sophisticated to handle complicated sentences such as ironic and satirical statements. Toward this end, combining multiple attention results from different perspectives on a sentence nonlinearly is preferred rather than relying on a single attention outcome. Besides, a holistic viewpoint to consider the entire context is required beyond the conventional word-focused thinking. For example, a simple sentiment classifier may report a negative sentiment on the battery in a review sentence. "Except for the battery, all other parts are not performing well". This is because it only concentrates on 'are not performing well'. To infer that the sentiment polarity toward the battery is in fact positive, our humans may first attend the phrase 'except for' and then 'are not performing well,' finally synthesizing multiple attention results from a global perspective reflecting the entire sentence.

Meanwhile, the above factors about location and content attention have been noticed by a few previous studies [34,35,37]. However, they have never been integrated all together. In addition, their respective underlying meaning does not seem to be well elucidated yet. Treating a sentence as a two-part object split with aspect terms, namely the target-dependent method, is known as if it is just for feeding target information into a deep learning model [37]. They do not appear to have indicated the directional position of relevant opinion words, i.e., left or right of the aspect terms. In addition, in a study that suggested nonlinear combination of several content attention results [34], the significance of considering a whole sentence was neglected. In another study that mentioned the importance of reflecting the entire sentence [35], other factors were ignored. In the present work, we merge respective factors based on contemplation of their usefulness with illumination of the meaning of relevant existing methods. Besides, we design two models differing in how to implement the target-dependent method and whether to incorporate the location attention. This differentiation plays a role in investigating relationships between model features and dataset characteristics raised in the second research question.

A review of previous studies suggests a tendency to seek a universal panacea which performs the best for all domains and languages regardless of dataset. None of these studies paid enough attention to the characteristics of datasets, although they often used the same benchmark datasets. However, knowledge of dataset will be a foundation for a deep understanding of the performance behaviors of a classifier. Further, such understanding can act as a groundwork for designing a better ABSA classifier reflecting domain characteristics for business application. As the first step in this direction, we suggest a new multifaceted perspective on the dataset, and then, together with the model factors identified from the first research question, investigate relationships between dataset characteristics and model features concerning classification performance. The new view on the dataset comprises data dimensions such as the number of aspects in a sentence, the sentimental heterogeneity of a sentence, the number of training samples per sequence of aspect terms, and the number of words in a preprocessed sentence. Harmonic mix of constituent factors may vary according to dataset characteristics represented by the data dimensions.

We perform experiments with three datasets widely used in ABSA research, laptop and restaurant reviews from SemEval 2014, and a twitter dataset from Dong et al. [45]. Our experimental results demonstrate that our models achieve state-of-the-art or comparable performances and that there are some useful relationships such as the suitability of bi-directional Long Short-Term Memory (LSTM) over one-directional LSTM for sentences containing multiple aspects and vice versa for sentences containing only one aspect. Our main contributions are as follows. First, we successively derive and integrate significant factors of an ABSA classifier, elucidating the meaning of relevant existing methods. Second, we suggest a new multilateral view on the dataset and start a systematic understanding of performance behaviors of a classifier. Third, we draw relationships between factors of model architecture and dimensions of dataset characteristics regarding classification performance. Concomitantly, we query a few fundamental issues that have been neglected and try somewhat different methods including evaluation scheme considering model stability. This work will provide considerable fresh insights into current ABSA research.

The remainder of the paper is organized as follows. Section 2 highlights previous research related to ABSA. Section 3 describes our proposed models in detail. Section 4 presents extensive experiments with discussion about their results. Section 5 concludes this work and suggests future research direction.

2. Related work

2.1. Studies on ABSA

A lot of studies have sought more intelligent methods for sentiment analysis due to its great potential for organizations and individuals [1–5]. Their focus naturally shifted from document or sentence to aspect level to obtain more valuable insights. The majority of studies on ABSA were devoted to four main tasks (i.e., aspect category detection, aspect term extraction, aspect sentiment classification, and aspect term sentiment classification) that can be grouped further according to task precedence or property: (1) aspect or aspect term identification; and (2) sentiment classification.

Previous studies on aspect or aspect term identification largely employed statistics-based, syntax-based, machine learning, or hybrid approach [46]. The statistics-based approach identifies aspects or aspect terms utilizing not only mere frequency, but also more sophisticated measures based on frequency such as pointwise mutual information, association rule, and co-occurrence matrix [47,48]. It tends to work well with high-frequency terms. However, it may fail in addressing infrequent terms. Meanwhile, the syntax-based approach exploits grammatical patterns and syntactical relationships often detected on a dependency tree [49,50]. Compared to the statistics-based approach, the syntax-based approach can also identify low-frequency cases. However, it is vulnerable to grammatical inaccuracy of input sentences. In addition, it requires rigorous manipulation for diverse situations. Next, the machine learning approach can be supervised or unsupervised. Many studies have used supervised learning such as Conditional Random Field (CRF) [51,52], Support Vector Machine (SVM) [53], and Naïve Bayes [54]. Recent deep learning models are also based on the supervised approach [9,28–30,32,33]. On the other hand, some works have adopted unsupervised approach such as Latent Dirichlet Allocation (LDA) topic modeling, although LDA may also be regarded as a type of statistical method [55,56]. Finally, the hybrid approach utilizes more than one of preceding approaches to gain particular synergistic benefit [57–59]. Ontology building method frequently uses both statistical and syntactical information to extract aspects or aspect terms [57,58]. Hybridization may be designed in a serial or parallel fashion.

Meanwhile, studies on sentiment classification have generally adopted knowledge-based or machine learning approach [4,46]. The knowledge-based approach may utilize one or more sentiment lexicons such as SentiWordNet and SenticNet to determine sentiment scores of individual words, aggregating these scores to calculate the sentiment polarity of the concerned aspect terms [60]. Furthermore, it may apply linguistic rules such as negation or domain ontology established based on statistical, grammatical, and lexical information [57,58]. By the way, many studies have utilized supervised machine learning techniques such as SVM [59,61] and Naïve Bayes [62]. Almost all recent deep learning models for sentiment classification rely on the supervised approach [31,32,34,35]. On the other hand, some works have employed the unsupervised approach such as LDA topic modeling for sentiment classification [55].

A main weakness of knowledge-based approach is that its validity relies heavily on the depth and breadth of mobilized resources [4]. Without an extensive and thorough knowledge base, it will be barely conceivable for a sentiment classifier to grasp most semantics implied in diverse natural language expressions. Another drawback of knowledge-based approach is the typicality of its knowledge representation. The strictly defined flat representation is likely to limit the handling of different affective features or nuances. For example, a sentiment lexicon 'unpredictable' can express positive sentiment when used like 'unpredictable plot', although it is defined as strong negative in SentiWordNet. Meanwhile, the traditional machine learning approach is labor-intensive mainly due to feature engineering which manually develops a set of significant features such as sentiment lexicons, bag-of-words, and syntactic patterns to enhance model performance and often acts as a performance bottleneck [25,26].

In contrast, a deep learning model automatically generates sentence feature representations from initial word vectors in an end-to-end manner. Besides, some knowledge sources such as sentiment lexicon, Part-Of-Speech (POS), and distance information can be appended to input word vectors to be refined automatically by a deep learning model. In this way, the spontaneous knowledge acquisition capability of deep learning is reshaping the research flow of ABSA. However, deep learning is not free from the weakness of supervised machine learning techniques, namely the requirement of a labeled dataset that is large enough. In this regard, there exists a favorable point in knowledge-based or unsupervised machine learning approach. Of course, unsupervised machine learning approach also has a few drawbacks. For example, topic modeling in ABSA suffers from inadequacy for fine-grained analysis, requirement of manual topic labeling, and relatively low classification accuracy.

This is a brief summary of the current research landscape of ABSA to ease understanding of our work in conjunction with previous studies. Many studies have also endeavored to improve the quality of word vectors used as input of a deep learning model to enhance the classification performance [63–70]. For a more comprehensive overview, refer to a few survey works [4,12,46]. Studies on attention mechanism will be discussed in the following section.

2.2. Attention mechanism

Attention mechanism assigns an importance weight to each lower level element when deriving an upper-level representation considering individual importance of element [19]. It was proposed in machine translation for the purpose of selecting referential words in original language for words in counterpart language before translation [19]. Afterwards, it has been applied in other domains such as question answering [21,22,71] and image caption generation [72]. In ABSA research, an attention mechanism has been utilized to compose a sentence representation with respect to given aspect terms. It realizes the intuition that not all words in a sentence equally contribute to the semantic meaning of the sentence and that the importance of a word should vary according to the aspect terms concerned. For example, in the above review sentence, “The food was good, but the service was terrible”., opinion word 'good' should be more emphasized than ‘terrible’ for aspect term ‘food’ whereas ‘terrible’ is more important for aspect term ‘service’.

Prior studies on ABSA have utilized two types of attention: location attention and content attention [34,36]. Location attention deals with positional relatedness of a word to a given aspect [34,36,42]. Intuitively, an opinion word closer to the given aspect terms should receive more importance than a farther one. Thus, location attention usually depends on the absolute distance between each word and the focal aspect terms in an input sequence. On the other hand, content attention evaluates semantic relatedness of each word to a given aspect. Three methods for implementing content attention have been reported. First, the inner product of each word and the aspect vector can model an element-wise interaction between the corresponding word and aspect vector. Second, the multiplication of a parameter weight matrix and the concatenated word-aspect vector allows a word with more likelihood to appear frequently with the aspect to get a higher weight [31,34,36]. Third, the multiplication of a word vector, a parameter weight matrix, and an aspect vector expresses more complex interaction among elements of the corresponding word and aspect vectors than the older inner product [44].

With only location attention, relevant opinion words far from the aspect terms can be almost ignored. On the other hand, with only content attention reflecting semantic relatedness between a given aspect and each context word [31,34,36,43,44], common opinion words that usually come with many different aspects may get similar weights. This can result in semantic mismatch problem between aspect terms and opinion words [35]. For example, in the above sentence “The food was good, but the service was terrible”., content attention weight for ‘terrible’ can be equal to or greater than that for ‘good’ when aspect ‘food’ is concerned. This is where the role of location attention gets significant. Therefore, a harmonious combination of these two attention mechanisms may be essential to have a useful sentiment classifier.

2.3. Deep learning models for aspect term sentiment classification

As aforementioned, this work focuses on aspect term sentiment classification with the supervised machine learning approach. Creative deep learning models have been actively proposed in this field recently. Target-Dependent LSTM (TD-LSTM) incorporates information of target, namely aspect terms of concern in a sentence, into the traditional standard LSTM [37]. It divides a sentence into left and right parts by the aspect terms and runs two parallel LSTM networks, namely left forward and right backward LSTM, concatenating the last hidden outputs to make sentence representation for sentiment classification. Target-Connection LSTM (TC-LSTM) extends TD-LSTM by concatenating each word input vector and aspect vector gained from averaging aspect term vectors. TD-LSTM and TC-LSTM have significantly boosted the classification accuracy in an empirical experiment without using any syntactic parser or external sentiment lexicons compared with the conventional SVM classifier or LSTM.

Attention-based LSTM with Aspect Embedding (ATAE-LSTM) combines a content attention mechanism with the conventional LSTM to concentrate on different parts of a sentence in response to different aspects [31]. Content attention mechanism differentiates the importance of word according to relatedness of the corresponding hidden word vector with the given aspect. To better utilize aspect information, the model concatenates the given aspect vector from separate aspect embedding to each word input vector. Both word and aspect embedding are optimized during training.

Deep Memory Network (DMN) model applies content attention mechanism distinctively on each word base rather than word phrase base of an LSTM-dependent model such as ATAE-LSTM [36]. DMN comprises multiple computational layers with shared parameters to successively yield more refined sentence representation toward an aspect. Each computational layer of DMN computes not only content attention considering the interdependence between a given aspect and each word, but also location attention reflecting the absolute distance between the aspect terms and each word.

Interactive Attention Networks (IAN) notes that many aspects consist of more than one aspect terms and tries to mitigate the limitation of frequently used aspect term averaging method of having one aspect vector. It interactively learns content attention of aspect terms as well as context words, combining two vectors acquired in parallel LSTM modules to compose a final sentence representation for classification [44].

Recurrent Attention on Memory (RAM) starts from observations that challenging sentence structures can hinder accurate classification where opinion words are enclosed in complicated phrase contexts, far from aspect terms, or dependent on each other but dispersed in a sentence [34]. A single attention model such as ATAE-LSTM or a linear combination model such as DMN might have difficulty addressing these complications [34]. As a result, RAM employs multiple attention mechanisms with nonlinear combination. More specifically, RAM first distills hidden memory slices from a bi-directional LSTM (BLSTM) with location attention. It then non-linearly combines multiple content attention outcomes from position-weighted memory with a Gated Recurrent Unit (GRU).

Content Attention Model (CAM) emphasizes the importance of global perspective to handle complex sentences such as ironical and sarcastic statements [35]. It adopts a sentence level content attention mechanism taking into account an intermediate sentence representation in addition to aspect and word vectors. Furthermore, it adds an intermediate sentence representation to content attention to derive an ultimate sentence representation. The intermediate sentence representation is produced, reflecting context attention weights calculated through parallel GRU networks and MLP networks.

3. Proposed models for aspect term sentiment classification

We introduce two aspect term sentiment classifiers: one implementing Holistic Recurrent content attention on Target-dependent memories from One-directional one-layered networks (HRT_One), the other from Bi-directional bi-layered networks (HRT_Bi). As shown in Figs. 1 and 2, these two models differ in how to build target-dependent memories and whether to incorporate location attention. HRT_One runs two one-directional one-layered LSTM networks to produce memories whereas HRT_Bi employs two bi-directional bi-layered LSTM networks. Besides, HRT_One synthesizes constituent memories without applying location attention whereas HRT_Bi applies location attention to such memories before integrating them. They will be explained in more detail in later sections.

3.1. Problem formulation

As aforementioned, this work focuses on aspect term sentiment classification. A sentence can describe more than one aspect, and an aspect can be expressed by an aspect term or a few consecutive aspect terms in a sentence. Let $S = \{s_1, \ldots, s_{a_1}, \ldots, s_{a_A}, \ldots, s_M\}$ denote an input sentence comprising M words and $S_a = \{s_{a_1}, \ldots, s_{a_A}\}$ represent a sequence of concerned aspect terms comprising A words. The goal is to predict sentiment polarity of sentence S toward aspect terms S_a.

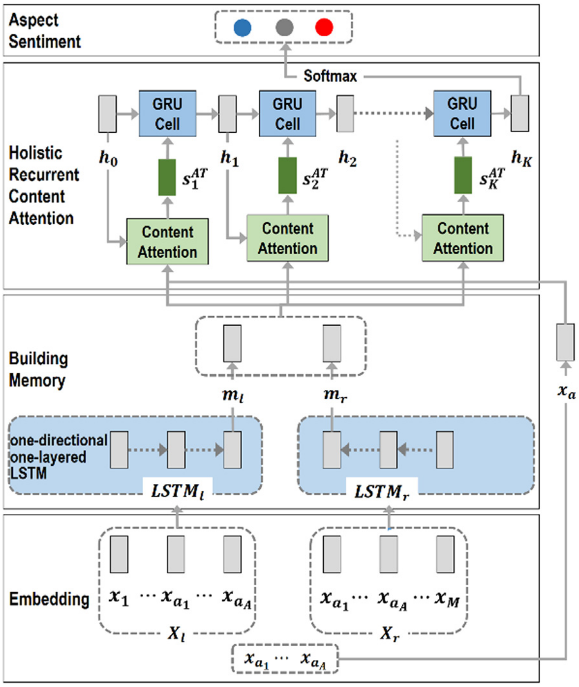

Fig. 1. Architecture of HRT_One.

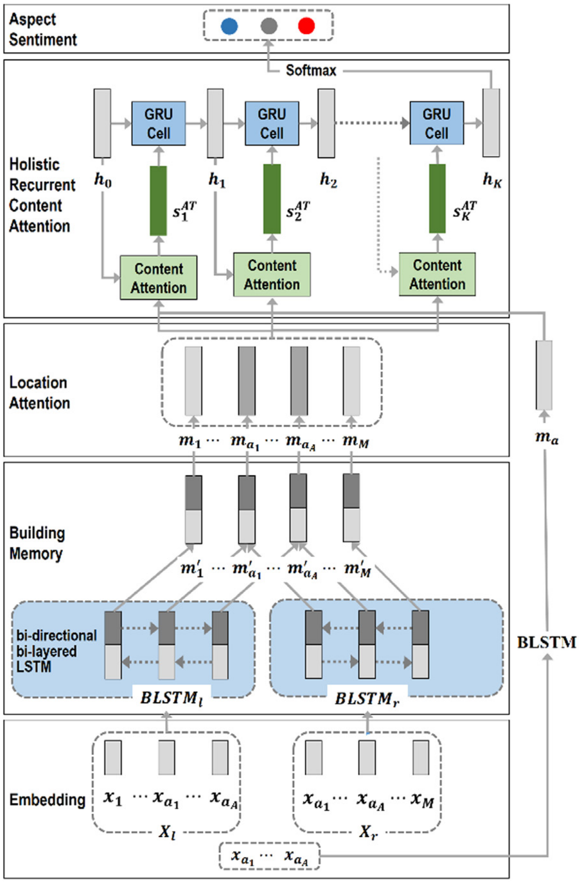

Fig. 2. Architecture of HRT_Bi.

3.2. Input embedding

Let us notate an embedding lookup table as $E \in R^{D\times|V|}$, where D is the dimension of word vectors and $|V|$ is vocabulary size. First, we split an input sentence S into the left part with aspect terms included in $S_l = \{s_1, \dots, s_{a_1}, \dots, s_{a_A}\}$ and the right part with aspect terms also included in $S_r = \{s_{a_1}, \dots, s_{a_A}, \dots, s_M\}$. Second, we convert each sentence part into a list of word vectors, namely, the left part $X_l = \{x_1, \dots, x_{a_1}, \dots, x_{a_A}\}$ and the right part $X_r = \{x_{a_1}, \dots, x_{a_A}, \dots, x_M\}$, where $x_t \in R^D$. A word vector x_t for a word s_t is obtained by $x_t = Eo_t$, where $o_t \in R^{|V|}$ is a one-hot vector of s_t. E can be configured to be constant during the training procedure to take advantage of the original semantics of input word vectors. Alternatively, it could be set to be tuned during the training procedure in order to capture some useful intrinsic information for sentiment classification.

3.3. Memory building

To build memory, we can basically employ LSTM to prevent gradient vanishing or exploding problem of conventional Recurrent Neural Networks (RNN). LSTM is good at capturing long-term dependencies and can easily handle sentences of any length [73, 74]. It has become a central architecture for sentiment analysis as well as other NLP tasks because its structure is naturally suited for many NLP applications [74]. We select one-directional and bi-directional LSTM networks to implement the target-dependent method and observe the difference of model behaviors according to dataset characteristics. Subsequently, we set the number of layers of HRT_One to one and that of HRT_Bi to two because each appears to be the best depending on exploratory experiments. For bi-directional case, two-layered structure has also been reported to perform well in NLP tasks [34,75].

3.3.1. HRT_One model

For our first model HRT_One, we will apply a forward LSTM to X_l and a backward LSTM to X_r and construct a composite memory $M = \{m_l, m_r\}$, where m_l is the last hidden state from the left forward LSTM and m_r is the last hidden state from the right backward LSTM.

3.3.2. HRT_Bi model

For each sentence part, we apply a bi-directional bi-layered LSTM in parallel, namely $BLSTM_l$ and $BLSTM_r$, and integrate the last outputs from each BLSTM into one sequence of memory. More specifically, at each time step t of layer z, the left forward LSTM $\overrightarrow{LSTM}_l$ takes in an element $\vec{h}_t^{(z-1)}$, an output from layer $z-1$ or an element x_t of X_l where z = 1 ($\vec{h}_t^0 = x_t$), stores a memory $\vec{c}_t^z$ inside its hidden memory cell, and produces hidden state $\vec{h}_t^z$ as follows:

$$\vec{i}_t = \sigma(\vec{W}_{ii}\vec{h}_t^{(z-1)} + \vec{b}_{ii} + \vec{W}_{hi}\vec{h}^z_{(t-1)} + \vec{b}_{hi})$$
$$\vec{f}_t = \sigma(\vec{W}_{if}\vec{h}_t^{(z-1)} + \vec{b}_{if} + \vec{W}_{hf}\vec{h}^z_{(t-1)} + \vec{b}_{hf})$$
$$\vec{g}_t = tanh(\vec{W}_{ig}\vec{h}_t^{(z-1)} + \vec{b}_{ig} + \vec{W}_{hg}\vec{h}^z_{(t-1)} + \vec{b}_{hg})$$
$$\vec{o}_t = \sigma(\vec{W}_{io}\vec{h}_t^{(z-1)} + \vec{b}_{io} + \vec{W}_{ho}\vec{h}^z_{(t-1)} + \vec{b}_{ho})$$
$$\vec{c}^z_t = \vec{f}_t \bigodot \vec{c}^z_{(t-1)} + \vec{i}_t \bigodot \vec{g}_t$$
$$\vec{h}^z_t = \vec{o}_t \bigodot \tanh(\vec{c}^z_t)$$

where $\vec{i}_t$, $\vec{f}_t$, $\vec{o}_t$ are input, forget, and output gates that control how much to take in the information from the current $\vec{g}_t$, how much to forget the information from the previous memory cell, and how much to reveal the information in the current memory cell as the output hidden state, respectively. σ and *tanh* represent sigmoid and hyperbolic tangent functions, respectively, and $\bigodot$ is an element-wise multiplication. Additionally, $\vec{W}_{ii}$, $\vec{W}_{if}$, $\vec{W}_{ig}$, $\vec{W}_{io}$ $\in R^{\vec{d}_z \times \vec{d}_{(z-1)}}$, $\vec{W}_{hi}$, $\vec{W}_{hf}$, $\vec{W}_{hg}$, $\vec{W}_{ho} \in R^{\vec{d}_z \times \vec{d}_z}$ are weight matrices, with $\vec{d}_z$ being the dimension of hidden states at layer z. Finally, $\vec{b}_{ii}$, $\vec{b}_{if}$, $\vec{b}_{ig}$, $\vec{b}_{io}$ and $\vec{b}_{hi}$, $\vec{b}_{hf}$, $\vec{b}_{hg}$, $\vec{b}_{ho}$ are all bias terms. Outputs generated from Z layers of $BLSTM_l$ are $M'_l = \{m'_{1,} \cdots, m'_{la_1}, \ldots, m'_{la_A}\}$, where $m'_t = \left(\overrightarrow{h^Z_t}, \overleftarrow{h^Z_t}\right) \in R^{\vec{d}_Z + \vec{d}_Z}$. As aforementioned, Z is set to two in our models. The bi-directional bi-layered LSTM for the right sentence part $BLSTM_r$ operates in the same manner as B$LSTM_l$ except that its input sequence X_r is reversed, yielding $M'_r = \{m'_{ra_1}, \ldots, m'_{ra_A}, \ldots, m'_M\}$. Combining M'_l and M'_r, we get averages of overlapped elements corresponding to aspect terms by $m'_{a_i} = (m'_{la_i} + m'_{ra_i})/2$ where $1 \leq \mathrm{i} \leq \mathrm{A}$, resulting in $M' = \{m'_{1,} \cdots, m'_{a_1}, \ldots, m'_{a_A}, \ldots, m'_M\}$.

3.4. Location attention: HRT_Bi model

Location attention is implemented straightforward based on the intuition that the closer an opinion word is to the aspect terms, the more it is likely to modify the aspect terms. Thus, to assign higher importance to the opinion word nearer to the aspect terms, we can make the location weight l_t for a word s_t vary according to the distance between an opinion word and the aspect terms. Specifically, the location weight l_t is calculated as follows:

$l_t = 1 - \frac{d_t}{n}$, where d_t represents distance, namely, $d_t = a_1 - t\,(1 \leq t < a_1)\,, d_t = 0\,(a_1 \leq t \leq a_A)\,, d_t = t - a_A\,(a_A < t \leq M)$,

Multiplying position weight to the corresponding hidden state output of the previous BLSTM gives the adjusted memory $\mathrm{M} = \{m_1,\ \cdots, m_{a_1}, \ldots, m_{a_A}, \ldots,\ m_M\}$, where $m_t = l_t \times m'_t \in R^{\vec{d}_Z + \vec{d}_Z}$. HRT_One model is designed not to contain a location attention module because including location attention during exploratory experiments does not appear to improve the performance. This may be partly attributed to the characteristic of one-directional LSTM that the latter word is considered more significantly in itself, resulting in relatively less benefit of location attention. By contrast, in a bi-directional LSTM such as HRT_Bi, a location attention appears to help because there is no natural gain from structural features.

3.5. Holistic recurrent content attention on memory

As mentioned in the Introduction, considering the whole sentence and integrating several different content attention outcomes nonlinearly will be significant factors for a useful ABSA classifier because it may acquire the ability of handling elusive sentences such as ironic and satirical statements [34,65]. Thus, we adopt GRU networks to nonlinearly combine intermediate sentence representations, which are obtained through different content attention weights on the memory reflecting the entire sentence. Note that the dimension of hidden memory from HRT_Bi is two times that from HRT_One because HRT_Bi uses bi-directional networks for memory building instead of one directional network in HRT_One. Thus, dimensions of parameter matrices should be adjusted accordingly. In addition, the process of deriving an aspect vector varies. This will be explained later in this section.

Now, the operation of this module, which is almost the same across the two models from a structural point of view, is as follows:

$$r_k = \sigma(W_{ir} s^{AT}_k + b_{ir} + W_{hr} h_{(k-1)} + b_{hr})$$

$$u_k = \sigma(W_{iu} s^{AT}_k + b_{iu} + W_{hu} h_{(k-1)} + b_{hu})$$

$$n_k = tanh(W_{in} s^{AT}_k + b_{in} + r_k \bigodot (W_{hn} h_{(k-1)} + b_{hn}))$$

$$h_k = (1 - u_k) \bigodot n_k + u_k \bigodot h_{(k-1)}$$

where h_k is the hidden state output at time k, where $1 \leq \mathrm{k} \leq \mathrm{K}$. The initial hidden state h_0 is a vector of 0. r_k and u_k are reset and update gates, respectively, that control the extent with which different information is blended. σ and $tanh$ represent sigmoid and hyperbolic tangent functions, respectively, with $\bigodot$ being an element-wise multiplication. W_{ir}, W_{iu}, $W_{in} \in R^{H \times \vec{d}_Z}$ are weight matrices in HRT_One and W_{ir}, W_{iu}, $W_{in} \in R^{H \times (\vec{d}_Z + \vec{d}_Z)}$ are weight matrices in HRT_Bi. W_{hr}, W_{hu}, $W_{hn} \in R^{H \times H}$ are weight matrices in both models, with H being the hidden size of GRU. s^{AT}_k is the input at time k. It is calculated considering content attention weight on each memory slice with the following equations.

$$c^k_t = W^{AT}_k (m_t; h_{(k-1)}; x_a)$$

$$\alpha^k_t = \frac{\exp(c^k_t)}{\sum_t \exp(c^k_t)}$$

$$s^{AT}_k = \sum_{t=1}^{M} \alpha^k_t m_t$$

In HRT_One, content attention weight at time k is computed reflecting the previous hidden state $h_{(k-1)}$ as well as the aspect vector $x_a = \sum_{i=1}^{A} x_{a_i}/A$. In contrast, x_a is substituted with m_a, resulting in $c^k_t = W^{AT}_k (m_t; h_{(k-1)}; m_a)$ in HRT_Bi. m_a is obtained by feeding given aspect terms to BLSTM networks and averaging their hidden output vectors. This is to make the dimension of m_a synchronized with other elements. This will also make aspect terms processed with a similar transformation as other context words. The previous hidden state $h_{(k-1)}$ implies the entire sentence information in this case because $h_{(k-1)}$ is an output to an input s^{AT}_{k-1} which is computed considering every memory slice from a sentence and can be considered as an intermediate sentence representation. Thus, this method adopts a holistic approach that is expected to be effective for complicated sentences.

3.6. Model training

The output vector from the last layer of the holistic recurrent content attention module is fed into a softmax layer for sentiment classification. The model is trained in a supervised manner to minimize cross-entropy loss given as follows.

$$\mathrm{L}\,(\theta) = -\sum_{(x,y)\in D} \sum_{c\in C} y^c \log f^c(x;\theta)$$

D denotes all training samples and C is the collection of sentiment classes. y^c is 1 or 0, indicating whether the ground truth class is c. $f^c\,(x;\theta)$ is the predicted probability for being sentiment class c given x and parameter set θ. We will employ Adam, an adaptive gradient algorithm, with weight decay regularization as an optimizer.

4. Empirical analysis

4.1. Datasets

We validate our proposed models with three benchmark datasets widely used in studies on ABSA. Two of them are laptop and restaurant reviews from SemEval 2014 and the third one is a twitter dataset collected by Dong et al. [45]. Table 1 shows statistics about sentiment labels of these datasets. A single review sentence is replicated for each sequence of aspect terms with information about aspect terms, their positions, and sentiment polarity attached. For example, for a restaurant review sentence, “The wait staff is friendly, and the food has gotten better and

Table 1
Statistics about the sentiment label of datasets.

Dataset		Pos.		Neg.		Neu.		Aspect sentence	
Laptop	Train	987	43%	866	37%	460	20%	2313	100%
	Test	341	53%	128	20%	169	26%	638	100%
Restaurant	Train	2164	60%	805	22%	633	18%	3602	100%
	Test	728	65%	196	18%	196	18%	1120	100%
Twitter	Train	1561	25%	1560	25%	3127	50%	6248	100%
	Test	173	25%	173	25%	346	50%	692	100%

Table 2
Statistics about the number of aspects in a sentence according to dataset.

Dataset		1 aspect	2 aspect	3 aspect	4 aspect	5 aspect	More aspect	Total sentence	1 Aspect Ratio
Laptop	Train	917	346	136	42	10	11	1462	63%
	Test	259	102	33	10	6	1	411	63%
Restaurant	Train	1008	556	260	102	32	20	1978	51%
	Test	285	192	73	31	14	5	600	48%
Twitter	Train	6240	1	0	0	0	0	6241	100%
	Test	692	0	0	0	0	0	692	100%

better!", two identical training aspect sentences are generated. One has aspect terms of 'wait staff' with positions from 4 to 14 and sentiment polarity of 'positive'. The other has aspect term of 'food' with positions from 36 to 40 and sentiment polarity of 'positive'. We will refer to the replicated sentence with a sequence of aspect terms as 'aspect sentence'. We will use 'a sequence of aspect terms' and 'aspect' interchangeably for simplicity. We exclude aspect sentence whose sentiment label is 'conflict', indicating mixed sentiment toward an aspect from the laptop and restaurant datasets. These aspect sentences are also removed in other studies because they can disturb the training process [34–36].

Most previous studies just presented data summary similar to Table 1, and did not explore the datasets further. However, it may be meaningful to investigate datasets from a multilateral perspective. Such investigation might lead to a deeper understanding of the performance behavior of a sentiment classifier. First of all, it will be worth taking a look at the overall distribution of the number of aspects in a sentence because if a sentence has more aspects, a sentiment classifier may encounter more difficulty during evaluation. Table 2 shows statistics about the number of aspects in a sentence. The number of aspect sentence in Table 1 includes duplicate sentences focusing on different aspects whereas the number of total sentences in Table 2 only counts single review sentences. The ratio of single review sentences describing only one aspect seems to be considerable. For example, this ratio in the twitter dataset is almost 100%. The fact that almost all of the sentences describe only one aspect may undermine the adequacy of the twitter dataset for ABSA. However, it may also allow us to gain more fundamental implications from simple sentence structure.

Second, it will be worth exploring the homogeneity of sentiment polarities of a sentence because an ABSA classifier is more likely to provide the truth if sentiment polarities are the same for all aspects in a sentence. Table 3 exhibits how many different review sentences express homogeneous or heterogeneous sentiment polarities toward different aspects. The 'Homo Sentiment' column counts the number of sentences with only positive, negative, or neutral labels while the 'Hetero Sentiment' column counts the number of other cases. We can see that the ratio of sentences expressing heterogeneous sentiment is around 10% for both laptop and restaurant datasets. In the twitter dataset where almost all sentences mention one aspect, the sentiment purity is almost 100% since the sentiment polarity toward one aspect is only one.

Third, the number of training samples per sequence of aspect terms and whether the aspect terms in test samples are in training samples will be worth exploring because a sentiment classifier may perform better toward the aspect terms for which it is trained with enough training samples. Table 4 shows that the number of training samples or test samples per aspect terms in the 'UAS/UAT' column is overwhelmingly higher in the twitter dataset than that in the other two datasets (UAS = Unique Aspect Sentence, UAT = Unique Aspect Terms). Incidentally, we found that there are some duplicate review sentences in the original training sets and in the training samples generated from them. Thus, we counted the number of aspect sentences from different review sentences. Besides, in the laptop and restaurant datasets, there are a considerable amount of test samples whose concerned aspect terms do not appear in training samples. Contrastively in the twitter dataset, the ratio of aspect terms that are included in test samples but not in training samples is much lower. A sentiment classifier may be trained and tested for each sequence of aspect terms to a higher extent with the twitter dataset than with the other datasets.

Fourth, word lengths of preprocessed input sentences that are fed into deep learning models will be worth investigating because longer input sentences may result in worse performance of a sentiment classifier. Table 5 illustrates that twits known to be limited to 40 words and regarded as short appear to be as long as other laptop or restaurant sentences after preprocessing.

4.2. Evaluation measures

We compare performances of different models using three evaluation metrics. The first is accuracy which is the proportion of true positives and true negatives to the total number of test samples examined. The second is macro-average F1 which is the average of F1 scores for each class. It has been regarded as reasonable when datasets exhibit an unbalanced class distribution as shown in Table 1 [34,37,45]. F1 score for each class is calculated as the harmonic mean of precision and recall of the class. However, the macro-average F1 score may not adequately reflect class imbalance because it takes the mean with equal importance for each class. Thus, the third metric, weighted-average F1 is employed. It is a weighted average of F1 scores for each class. Weights are proportional to the support, namely the number of true samples in each class.

Table 3
Statistics about the sentiment purity of a sentence in different datasets.

Dataset		Homo Sentiment		Hetero Sentiment		Total Sentence		Aspect Sentence	Aspect per Sentence
Laptop	Train	1297	89%	165	11%	1462	100%	2313	1.6
	Test	373	91%	38	9%	411	100%	638	1.6
Restaurant	Train	1659	84%	319	16%	1978	100%	3602	1.8
	Test	520	87%	80	13%	600	100%	1120	1.9
Twitter	Train	6241	100%	0	0%	6241	100%	6248	1.0
	Test	692	100%	0	0%	692	100%	692	1.0

Table 4
Statistics about the number of training samples per sequence of aspect terms of datasets.

Dataset		Aspect Sentence	Unique Aspect Sentence	Unique Aspect Terms	UAS/ UAT	Asp. Terms in Test but not in Train	Percent. of Asp. Terms in Test but not in Train
Laptop	Train	2313	2293	943	2.43	235	235/389 = 60%
	Test	638	638	389	1.64		
Restaurant	Train	3602	3589	1195	3.00	333	333/520 = 64%
	Test	1120	1120	520	2.15		
Twitter	Train	6248	6242	113	55.24	5	5/82 = 6%
	Test	692	692	82	8.44		

Table 5
Statistics about the number of words in a sentence after preprocessing of datasets.

Dataset		Avg.	Std.	Med.	Max.	Min.	Aspect Sentence
Laptop	Train	20.93	11.66	19	83	2	2313
	Test	17.27	10.81	15	71	2	638
Restaurant	Train	19.28	10.31	17	79	2	3602
	Test	18.19	9.62	16	70	3	1120
Twitter	Train	20.10	7.45	21	73	3	6248
	Test	20.26	7.35	21	40	3	692

4.3. Experimental setting

In addition to our proposed models, we implemented TD-LSTM and RAM which formed the basis of our models, reflecting their core ideas and tuning hyperparameters. TC-LSTM connects a concerned aspect vector obtained through averaging aspect term vectors to each input word vector. It has been reported that TC-LSTM performs worse than TD-LSTM in the aspect sentiment classification for the restaurant dataset [31]. We obtained similar results from exploratory experiments in aspect term sentiment classification. Thus, we chose TD-LSTM rather than TC-LSTM. We set the number of different content attentions, namely K in Section 3.5, as five in HRT_One, HRT_Bi, and RAM based on results from exploratory experiments.

Besides, hyperparameters unified across all four models are as follows. The embedding lookup table was built using the 42B token version of GloVe 300-dimensional word vectors [76]. All word vectors for vocabularies that are present in the datasets but not in GloVe were initialized with a uniform distribution U (−0.25, 0.25). Table 6 shows the number of total vocabularies in each dataset, the number of vocabularies matched with GloVe word vectors, and the percentage of matched vocabularies to total vocabularies. Input word vectors were set to stay static during the training procedure depending on results from exploratory experiments.

A dropout rate of 50% was applied to input word vectors to prevent overfitting. The batch size was 64. All trainable parameters were initialized with Xavier Uniform initializer. Adam optimizer was applied with learning rate 0.001 and weight decay regularization 1.0×10^{-5}.

Table 6
Statistics about vocabularies of datasets.

Dataset	Vocabulary (A)	Match. Vocab. (B)	Ratio (B/A)
Laptop	3,895	3,449	88.5%
Restaurant	4,953	4,302	86.9%
Twitter	16,096	11,501	71.5%

With these conditions, we carried experiments in two steps. In the first preliminary experiment, four models were compared with other previous models that used the three datasets mentioned in Section 4.1 in a somewhat rough manner. The maximum accuracy was recorded after one-time ten rounds of execution with each round comprising 40 epochs. Macro-average F1 was measured at the epoch with the maximum accuracy. Performances of all previous models were retrieved from corresponding papers. This comparison may be meaningful in some respect because each model must have been optimized for its own and share these datasets. In the second main experiment, four models were compared with more depth considering random factors inherent in the feeding order randomization of training data samples and random initialization of weight matrices. Performance metrics were derived in an average-based manner rather than in a one-time manner to handle arbitrary effects. The average of maximum accuracy of each round was recorded after 30 rounds of execution, with each round also comprising 40 epochs. In each round, weighted-average F1 and macro-average F1 were measured at the epoch whose accuracy was the maximum. Average values of all rounds were then obtained.

Table 7
Preliminary experimental results.

Model	Laptop		Restaurant		Twitter	
	Acc.	Mac. F1	Acc.	Mac. F1	Acc.	Mac. F1
SVM	(70.49)	(NA)	(80.16)	(NA)	(NA)	(NA)
TD-LSTM	72.57 (NA)	67.87 (NA)	80.09 (NA)	71.28 (NA)	73.27 (70.8)	71.48 (69.0)
TC-LSTM	(NA)	(NA)	(NA)	(NA)	(71.5)	(69.5)
ATAE-LSTM	(68.7)	(NA)	(77.2)	(NA)	(NA)	(NA)
IAN	(72.1)	(NA)	(78.6)	(NA)	(NA)	(NA)
DMN	(72.37)	(NA)	(80.95)	(NA)	(NA)	(NA)
RAM	**74.61** (74.49)	71.02 (71.35)	80.71 (80.23)	72.30 (70.80)	71.97 (69.36)	70.27 (67.30)
CAM	(75.07)	(NA)	(80.89)	(NA)	(71.53)	(NA)
HRT_One	73.04	68.69	81.43	72.57	**73.84**	**72.08**
HRT_Bi	74.45	70.83	**81.96**	**74.09**	73.27	71.98

Results in parenthesis are retrieved from corresponding papers. Best scores are in bold except for the accuracy of the laptop dataset. Refer to the manuscript for reasons.

4.4. Preliminary results

Table 7 shows preliminary experimental results of our models with other models' performances retrieved from original papers. Compared models are TD-LSTM [37], TC-LSTM [37], ATAE-LSTM [31], IAN [44], DMN [36], RAM [34], and CAM [35]. These were mentioned earlier in Section 2.3. The result of the best team in SemEval 2014 is based on the conventional state-of-the-art method SVM utilizing lexicon, parsing, and n-gram features [77]. It is added to help grasp overall performances of models. Figures in parenthesis show results reported in the corresponding paper, with 'NA' indicating not applicable (the work did not evaluate the metric). On the other hand, figures without parenthesis represent performances of models implemented in this work. TD-LSTM and RAM reproduced with hyperparameter tuning in the present work provide higher or similar level of performances compared to those in the original papers.

Overall, the best model appears to vary according to dataset. For laptop, restaurant, and twitter datasets, RAM, HRT_Bi, and HRT_One achieve the highest score, respectively, regarding both accuracy and macro-average F1 measures. The accuracy of CAM [35] has been reported to be higher than that of RAM for the laptop dataset. However, we focus on RAM. One of the reasons is that both HRT_Bi and HRT_One are more similar to RAM than CAM from a structural viewpoint and comparing with RAM will be more conducive to discover relations between dataset characteristics and model architecture. Simultaneously, in our understanding, RAM also reflects the entire sentence when computing content attention, and thereby the key idea of CAM. In addition, both HRT_Bi and HRT_One seem to perform better than CAM in the restaurant and twitter datasets. Thus, we conduct a more in-depth analysis with implemented models: TD-LSTM, RAM, HRT_One, and HRT_Bi.

4.5. Main results

Table 8 presents 30-round results from four implemented models with their averages and sample standard errors. A trend similar to that found in Table 7 is also observed regarding the three metrics including weighted-average F1. For laptop, restaurant, and twitter datasets, RAM, HRT_Bi, and HRT_One appear to be the best, respectively. As expected, however, specific average scores shown in Table 8 are overall less than those shown in Table 7 where the maximum values are recorded. In the restaurant dataset, we can observe that the average score of HRT_Bi in Table 8 is higher than the maximum value of CAM in Table 7. Similar case is found for HTHR_One against CAM in the twitter dataset.

To examine relationships inherent in these results more thoroughly, we conduct pair-wise hypothesis testing concerning average and variance differences. The average of 30-round values of each metric naturally imply the level of model performance. The variance of them may well be an indicator of model stability. Table 9 summarizes results of the two-sample Z- and F-tests for average and variance differences, respectively, by a one-tailed method. An additional two-sample T-test for average difference gave the same results as those from the two-sample Z-test. Model names in cells where the same model is better than the other in both average and variance with at least one metric statistically significant are marked in italics and color.

In the laptop dataset, HRT_Bi seems to be significantly better and more stable than TD-LSTM, with RAM being significantly better than HRT_Bi in average performance but not confident of stability. In the restaurant dataset, HRT_Bi appears to perform better and more reliably than TD-LSTM or RAM. In the twitter dataset, HRT_One shows higher and more reliable performance than TD-LSTM or RAM, with TD-LSTM being better than

Table 8
Results of 30-round execution.

Model	Laptop			Restaurant			Twitter		
	Acc.	Mac. F1	Wt. F1	Acc.	Mac. F1	Wt. F1	Acc.	Mac. F1	Wt. F1
TD-LSTM	0.7138 (0.0094)	0.6565 (0.0127)	0.7107 (0.0111)	0.7988 (0.0038)	0.7003 (0.0102)	0.7876 (0.0056)	0.7246 (0.0080)	0.7038 (0.0114)	0.7214 (0.0092)
RAM	**0.7359** (0.0073)	**0.6895** (0.0116)	**0.7358** (0.0091)	0.7998 (0.0065)	0.6975 (0.0157)	0.7878 (0.0088)	0.7150 (0.0057)	0.6959 (0.0074)	0.7125 (0.0063)
HRT_One	0.7155 (0.0054)	0.6587 (0.0087)	0.7115 (0.0073)	0.7985 (0.0044)	0.6984 (0.0119)	0.7872 (0.0069)	**0.7250** (0.0054)	**0.7061** (0.0079)	**0.7225** (0.0061)
HRT_Bi	0.7310 (0.0051)	0.6783 (0.0107)	0.7274 (0.0079)	**0.8119** (0.0040)	**0.7214** (0.0082)	**0.8034** (0.0049)	0.7186 (0.0074)	0.6971 (0.0105)	0.7152 (0.0083)

Figures in parenthesis represent samples' standard errors. Best scores are in shown bold.

Table 9
Results of average and variance difference test.

Data	Mod.	Test Item	Acc.		Mac. F1		Wt. F1	
			TD-LSTM	RAM	TD-LSTM	RAM	TD-LSTM	RAM
Lapt.	HRT_ One	Avg.	_HRT_One_	RAM***	_HRT_One_	RAM***	_HRT_One_	RAM***
		Var.	_HRT_One***_	HRT_One**	_HRT_One**_	HRT_One*	_HRT_One**_	HRT_One
	HRT_ Bi	Avg.	_HRT_Bi***_	_RAM***_	_HRT_Bi***_	RAM***	_HRT_Bi***_	RAM***
		Var.	_HRT_Bi**_	_RAM_	_HRT_Bi_	HRT_Bi	_HRT_Bi**_	HRT_Bi
Rest.	HRT_ One	Avg.	TD-LSTM	RAM	TD-LSTM	HRT_One	TD-LSTM	RAM
		Var.	TD-LSTM	HRT_One**	TD-LSTM	HRT_One*	TD-LSTM	HRT_One*
	HRT_ Bi	Avg.	HRT_Bi***	_HRT_Bi***_	_HRT_Bi***_	_HRT_Bi***_	_HRT_Bi***_	_HRT_Bi***_
		Var.	TD-LSTM	_HRT_Bi_	_HRT_Bi_	_HRT_Bi***_	_HRT_Bi_	_HRT_Bi***_
Twit.	HRT_ One	Avg.	_HRT_One_	_HRT_One***_	_HRT_One_	HRT_One***	_HRT_One_	_HRT_One***_
		Var.	_HRT_One**_	_HRT_One_	_HRT_One**_	RAM	_HRT_One**_	_HRT_One_
	HRT_ Bi	Avg.	TD-LSTM***	HRT_Bi**	TD-LSTM**	HRT_Bi	TD-LSTM***	HRT_Bi*
		Var.	HRT_Bi	RAM*	HRT_Bi	RAM**	HRT_Bi	RAM*

Model names in cells where a model is better than the other in both average and variance with at least one metric statistically significant are shown in italics with blue- or green-color according to the number of significant metrics.
*: p-v. < 0.1, **: p-v. < 0.05, ***: p-v. < 0.01.

HRT_Bi and HRT_Bi being better than RAM regarding average performances.

It has been reported that TD-LSTM equipped with attention mechanism does not show any improvement [37]. It is not so effective either because it cannot know which words are significant for a given aspect [31]. In addition, it might lose the sentiment feature of opinion words when they are far from given aspect terms [34,35]. Expectedly, TD-LSTM performs invariably less than RAM or HRT_Bi in the laptop and restaurant datasets. However, TD-LSTM tends to surpass RAM and HRT_Bi in the twitter dataset. To understand what makes these differences might be of more value than to merely compare performances of models. In this respective, relationships between the architecture of models and characteristics of datasets need to be examined regarding performance results.

Distinctive features of the twitter dataset compared to other datasets may be summarized as follows according to the statistics presented in Section 4.1. The number of aspects per sentence is mostly one and the sentiment polarity of a sentence is usually pure. Additionally, the number of training samples per aspect is much higher than those of the other datasets and the test samples contain almost the same aspect terms that the classifier is trained for.

In contrast, the laptop and restaurant datasets have many similarities. The number of aspects per sentence is greater than one, ranging from 1.6 to 1.9. The ratio of sentences with heterogeneous sentiment polarity is about 10%, ranging from 9% to 16%. Additionally, the number of training samples per sequence of aspect terms is very small, ranging from 2.43 to 3.00. The ratio of aspect terms that are in test samples but are not in training samples is considerable, ranging from 60% to 64%

Considering all things synthetically, we can observe a few interesting relationships. When the dataset comprises sentences with only one aspect as in the twitter dataset, target-dependent one-directional LSTM architecture such as HRT_One or TD-LSTM seems to be a useful alternative. A more complex architecture such as HRT_Bi or RAM tends to yield rather worse outputs. These results may be ensured when the number of training samples per aspect is high enough for a target-dependent one-directional model to learn semantic features for each aspect as in the twitter dataset.

On the other hand, when the dataset consists of sentences with at least one aspect as in the laptop or restaurant dataset, a bi-directional LSTM architecture with location attention such as HRT_Bi or RAM appears to be a reliable alternative. If multiple word vectors pertaining to different aspects are blended in target-dependent one-directional way, the ultimate sentence representation may not be so useful as in the one-aspect sentence case. When there are multiple aspects in a sentence, the architecture to elicit richer semantic information from both directions seems to be more effective even with insufficient training samples. Besides, location attention seems to play a significant role in this situation. Comparing target-dependent bi-directional architecture such as HRT_Bi with target-independent bi-directional architecture such as RAM, it seems that the former is better for the multiple-aspect case. However, results are not so apparent in the laptop dataset as in the restaurant dataset. This phenomenon may be attributed to differences in the number of aspects in each domain and the number of parameters to be trained in each model. There are more aspects in the laptop domain than those in the restaurant domain whereas there are more parameters in HRT_Bi than those in RAM. Thus, HRT_Bi may become weaker

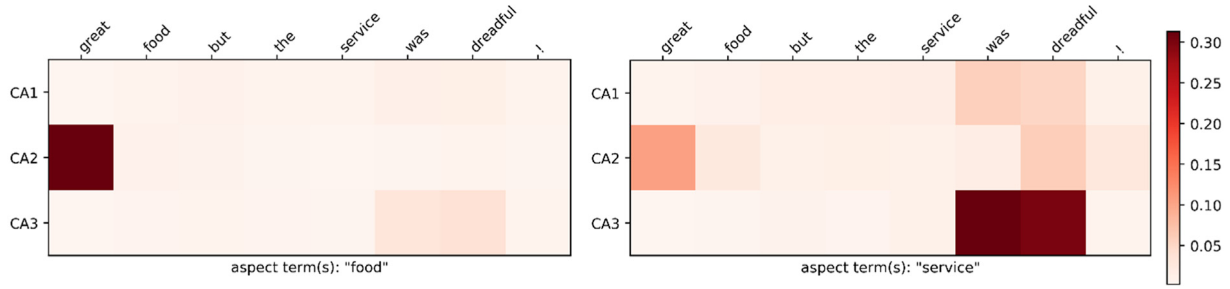

Fig. 3. Example of content attention on a heterogeneous sentence by HRT_Bi.

than RAM against the laptop dataset, although its stability does not deteriorate so much.

Regarding aspects of each domain, those of restaurant are set as food, service, price, ambiance, and anecdotes/miscellaneous in SemEval 2014 whereas those of laptop are not set explicitly in SemEval 2014. However, they can be conjectured based on the dataset for aspect term sentiment classification to be price, size, weight, screen quality, memory, keyboard, CPU, battery, storage, USB ports, durability, built-in apps, case, customer service, mouse, OS, touchpad, and so on. The smaller the number of aspects, the smaller the sentence variations will be, although there can exist different aspect terms for an aspect. Therefore, sentence variations originating from the variety of opinion words or semantic expressions for each aspect are likely to be greater in the laptop dataset than those in the restaurant dataset.

Meanwhile, we may well expect better results from the twitter dataset where most sentences mention one aspect with homogeneous sentiment polarity than those from the other datasets. The twitter dataset also seemed to be more promising when the number of training samples per sequence of aspect terms is compared. However, results are worse than those from restaurant dataset commonly across the four models as shown in Tables 7 and 8. Probable reasons may be the outstanding number of vocabulary and the far lower ratio of matched word vectors as shown in Table 6. Another reason may be higher expressiveness, more sarcasms, less grammatical correctness compared to review sentences [78] which cannot be measured easily for now. It needs to be measured in future research.

4.6. Results for sentences expressing heterogeneous sentiment polarities

If a test sample is homogenous in sentiment polarity toward all aspects in the sentence, an ABSA classifier may seem to perform well although its attention mechanism does not work effectively. In other words, an ABSA classifier with a good attention mechanism will yield a similar level of performance, whether evaluated sentences are homogeneous or not in sentiment polarity. Previous studies do not seem to have taken any notice of this point. They tend to show only an illustrative case example exhibiting a weight distribution over each word in a sentence. We prefer to employ a more macro approach and check the average of classification accuracy for sentences with heterogeneous sentiment polarities (Acc_Het_Senti) as shown in Table 10. Average values were obtained similarly as those for metrics shown in Table 8 through 30-round execution.

As expected, the accuracy of sentences with heterogeneous sentiment polarities is much lower than that of all test samples commonly across the four models. For the laptop test set, the accuracy difference ranges from 12.6% to 14.0%. For the restaurant dataset, it ranges from 13.8% to 16.3%. RAM seems to perform better than TD-LSTM or HRT_One in the restaurant dataset. However, for heterogeneous sentences of the dataset, the other two models appear to outperform RAM. The best performing HRT_Bi is also the most robust against heterogeneous samples. Consequently, the results from the restaurant dataset appear to support the advantage of target-dependent method for heterogeneous sentences as well as for all test samples. Although this phenomenon was not so apparent in the laptop dataset as in the restaurant dataset, accuracy differences between RAM and TD-LSTM or between RAM and HRT_One are smaller for heterogeneous sentences than those for all test samples.

These results may need some caution to be generalized due to the far smaller proportion of sentences with heterogeneous sentiment polarity as shown in Table 3. However, we cannot deny the fact that all four models use the same datasets and that a classifier with a perfect attention mechanism and architectural robustness will not be affected by sentimental heterogeneity of a sentence.

4.7. Qualitative analysis

In addition to macro observation shown in Section 4.6, we take a micro approach to investigate how holistic recurrent content attention mechanism might work. Although content attention results are adjusted further through element-wise operations of GRU, an attention plot will help us grasp and check the overall logic. Fig. 3 shows three different layers of content attention weights corresponding to α_t^k when $1 \leq k \leq 3$ in Section 3.5 of HRT_Bi model. Color depth indicates the magnitude of content attention, with darker color meaning higher weight and greater importance. The example sentence is "Great food but the service was dreadful!". The sentiment polarity toward "food" is positive but that toward "service" is negative. Three different perspectives on a sentence seem to enable more reliable derivation of sentiment polarity. For example, without the second content attention layer for "food" where 'great' has remarkable importance, the sentiment polarity would be classified as negative because 'dreadful' has a slightly higher weight than the other words in the first and third layers. On the other hand, with only the second content attention layer for "service" where 'great' is more emphasized than 'dreadful', the sentiment polarity would be mistakenly classified as positive. Considering the first and third attention layers together with the second one, HRT_Bi appears to draw the truth.

In Fig. 4, the example sentence is "It is very fast and has everything that I need except for a word program". HRT_Bi seems to attend 'except for' at the first and second attention and then 'very fast and has everything', drawing negative sentiment polarity for "word program" after synthesizing all information.

4.8. Error analysis

Our general error analysis of the HRT_Bi model on the restaurant dataset found that many misclassified sentences have one of the following forms. The first is an indirect expression, such as

Table 10
Accuracy of sentences with heterogeneous sentiment polarities from 30-round execution.

Model	Laptop			Restaurant		
	Accuracy (A)	Acc_Het_Senti (B)	Difference (A−B)	Accuracy (A)	Acc_Het_Senti (B)	Difference (A−B)
TD-LSTM	0.7138	0.5874	0.1264	0.7988	0.6499	0.1489
RAM	0.7359	0.5997	0.1362	0.7998	0.6373	0.1625
HRT_One	0.7155	0.5851	0.1304	0.7985	0.6463	0.1522
HRT_Bi	0.7310	0.5906	0.1404	0.8119	0.6735	0.1384

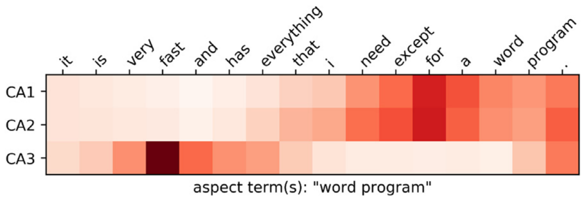

Fig. 4. Example of content attention on a complicated sentence by HRT_Bi.

satire, which often involves comparison, negation, interrogation, etc. A sentence of the test set, "Frankly, the chinese food here is something I can make better at home", expresses a negative sentiment toward 'chinese food' through comparison. For easy recognition, the concerned aspect terms are underlined. Another sentence, "Our waiter was friendly, and it is a shame that he didn't have supportive staff to work with", describes 'staff' as negative through negation. Additionally, an interrogative sentence, "How can they hope to stay in business with service like this?" mentions 'service' with a negative feeling. The second mistakable form contains a sentimental expression that consists of two or more words, which has a different meaning from the original words. In the sentence, "The sangria's watered down", 'watered' and 'down' together express a negative sentiment. The third is an unusual sentimental word that expresses different sentimental polarities according to the corresponding aspect terms. In the sentence, "The steak melted in my mouth", the sentimental word 'melted' expresses a positive sentiment. However, a sentence of the training set, "Butter was melted, white wine warm, and cheese oozing everywhere", has a negative label.

All these errors will be closely related to the number of training samples for each sentence pattern. With the lack of corresponding training samples, the HRT_Bi model may not act properly despite being equipped with holistic recurrent attention. Thus, a more thorough analysis would consider the number of training samples for each sentence pattern with respect to each unique sequence of aspect terms in the test set. However, this kind of work will require a considerable amount of effort, due to the diversity of language expression, and is omitted in this study.

4.9. Other findings and issues

Attention is usually paid to each word vector that represents the semantics of each word in the memory network that composes a sentence representation through a summation of weighted word vectors [36]. In contrast, attention is mostly paid to each hidden vector that contains the semantics of preceding words in a recurrent network [31,34,44]. However, attention does not always appear to guarantee performance gains. HRT_Bi model that adopts two bi-directional LSTM networks achieved a slightly higher performance, when incorporating location attention for the laptop and restaurant datasets. However, this was not the case for the twitter dataset where only one aspect is mentioned in one sentence in most cases. In the twitter dataset, the location attention of HRT_Bi tended to rather degrade the performance. Besides, in the HRT_One or TD-LSTM model that employ two one-directional LSTM networks, no gains were observed from location attention, similar to the results in other studies [37]. This phenomenon may be partly because in the one-directional LSTM network, the latter word that is closer to the target naturally has higher importance. As for RAM that does not employ the target-dependent method, location attention appeared to be slightly conducive to the classification performance for all three datasets. In summary, in LSTM networks that are bi-directional without employing the target-dependent method or when LSTM networks are applied to sentences with more than one aspect, location attention is likely to be conducive to the performance of a classifier. In contrast, in LSTM networks that are one-directional, or adopt target-dependent method, or when LSTM networks are applied to sentences whose aspect number is one, location attention might be rather harmful.

On the other hand, content attention used in HRT_Bi, RAM, and HRT_One for deriving a sentence representation as an input to GRU cell appears to play an assistive role. Here, content attention evaluates the semantic importance of each memory reflecting the concerned aspect and whole sentence. GRU cell that nonlinearly combines several different sentence representations generated from sequentially updated content attention can produce a more suitable sentence representation. In contrast, TD-LSTM does not seem to gain any benefit from content attention considering the relationship between each memory slice and given aspect.

HRT_Bi may take longer than RAM to be trained because the former has more parameters than the latter. Once it is trained, however, its speed of computing new sentences through the model will not matter anymore. In other words, there will be little difference in the time needed for any model to classify new sentences in a production environment. Therefore, training time may be something not to seek too rigorously, but to meet within an acceptable range in most business applications. The emphasis may be placed on the performance of a sentiment classifier rather than on the training time if both could not be fulfilled. In this respect, various network architectures employing LSTM may need be explored, although they are a little slower in training than those based on memory network or Convolutional Neural Network (CNN).

In this work, we measured the average performance and stability of each model through 30 rounds of execution, each round comprising 40 epochs, as shown in Table 8. A general one-time method of performance measurement utilizing a validation set and an early stopping were not adopted to mitigate effects of arbitrary and random factors. Average scores of Table 8 are overall lower than those of Table 7 evaluated in a one-time manner which is more similar to the general method. Checking both average performance and stability of a model seems to be more reasonable than only focusing on the maximal performance, especially when random factors affect the performance and when even a little difference in performance matters. The development of GPU is expected to enable a more objective and systematic performance measurement.

5. Conclusions and future work

We proposed two models employing target-dependent LSTM to produce memories from an input sentence and GRU cells for integrating sentence representations generated from different content attention weights on memories. HRT_Bi appears to achieve state-of-the-art performance in the restaurant dataset and comparable performance in the laptop dataset, with HRT_One accomplishing the highest performance in the twitter dataset. Based on these results, we have come to reilluminate the usefulness of target-dependent method which has been gradually regarded obsolete by more recent studies. In addition, we demonstrated the effectiveness of integrating target-dependent method and holistic recurrent content attention. Besides, by turning our attention to datasets, we derived some interesting relationships between characteristics of datasets and model architecture. One of such relationships is that one-directional networks as in HRT_One and TD-LSTM models are more effective than bi-directional networks as in HRT_Bi and RAM for sentences such as those in the twitter dataset. However, this situation is reversed for sentences such as those in the laptop and restaurant datasets.

Our future ABSA research may start from well-established datasets. These datasets will include enough training and test samples for each aspect or each sequence of aspect terms with sufficient homogeneous and heterogeneous samples, one-aspect samples, and multi-aspect samples. The construction of datasets will be of great importance for business applications as well as for academic purpose, although it may involve substantial costs. On the other hand, studies on how to build desirable datasets with little costs or how to train a classifier with small datasets may be required.

With adequate balanced datasets, a hybrid approach that applies two different models according to the number of aspects mentioned in a sentence may be worth exploring. More specifically, we can apply a model such as HRT_One trained with one-aspect sentences for one-aspect test sentences while we can use a model such as HRT_Bi trained with one-aspect and multi-aspect sentences for multi-aspect test sentences. We could not carry out this type of experiment with the benchmark laptop and restaurant datasets due to their small sizes. According to results of the present work, when the ratio of one-aspect sentences is considerably high, a model such as HRT_One appears to perform better for these sentences. These results of course need to be confirmed through more empirical validations.

One more thing to contemplate is that there exist a noticeable amount of multi-aspect sentences including both implicit and explicit aspects. However, recent ABSA studies tend to pay much more attention to aspect term classification task dealing with only explicit aspects. The technique used in TC-LSTM and ATAE-LSTM to concatenate an aspect vector to each input word vector for aspect term classification has been reported to be worse than that used in more recent studies without concatenation. Additionally, the current location attention mechanism is almost impossible to apply to an implicit aspect because its position is not exposed. There are cases where an implicit aspect is left out while only explicit aspect terms are labeled or several groups of aspect terms in a sentence correspond to only one aspect in the restaurant dataset of SemEval 2014. Therefore, a smarter model for aspect sentiment classification to address both implicit and explicit aspects is warranted. One experimental alternative is aspect embedding to allow an ABSA classifier to learn aspect vectors because an implicit aspect can also have an aspect vector in this way. A more advanced alternative coupled with aspect embedding may be infusing aspect location information into neural networks to compensate for the absence of location attention. Aspect location information whose dimension equals to the number of aspects in the corresponding domain can be acquired from co-occurrence values of each aspect and word pairs contained in one aspect sentences and then concatenated to input word vectors.

Lastly, a more intelligent mechanism that is robust over sentences with heterogeneous sentiment polarities needs to be created. We revealed that classification performances of all models implemented in this work were far lower for these sentences than those for homogeneous sentences. Although the target-dependent method seemed to be a little stronger than the unified approach, further research should be devoted to this topic.

Acknowledgments

This work was supported by the Ministry of Education, Republic of Korea. It was also supported by a grant (NRF-2017S1A5B5A02024287) of the National Research Foundation of Korea. We thank anonymous reviewers for their useful comments and suggestions.

References

[1] O. Appel, F. Chiclanaa, J. Carter, H. Fujita, A consensus approach to the sentiment analysis problem driven by support-based IOWA majority, Int. J. Intell. Syst. 32 (9) (2017) 947–965.
[2] O. Appel, F. Chiclanaa, J. Carter, H. Fujita, Cross-ratio uninorms as an effective aggregation mechanism in sentiment analysis, Knowl.-Based Syst. 12 (2017) 16–22.
[3] O. Appel, F. Chiclanaa, J. Carter, H. Fujita, Successes and challenges in developing a hybrid approach to sentiment analysis, Appl. Intell. 48 (5) (2018) 1176–1188.
[4] E. Cambria, Affective computing and sentiment analysis, IEEE Intell. Syst. 31 (2) (2016) 102–107.
[5] N. Dosoula, R. Griep, Rick den Ridder, R. Slangen, Ruud van Luijk, K. Schouten, F. Frasincar, Sentiment analysis of multiple implicit features per sentence in consumer review data, in: Proceedings of the 12th International Baltic Conference on Databases and Information Systems, 2016, pp. 241–254.
[6] O. Araque, G. Zhu, C.A. Iglesias, A semantic similarity-based perspective of affect lexicons for sentiment analysis, Knowl.-Based Syst. 165 (2019) 346–359.
[7] B. Liu, Sentiment analysis and opinion mining, Synth. Lect. Hum. Lang. Technol. 5 (1) (2012) 1–167.
[8] B. Pang, L. Lee, Opinion mining and sentiment analysis, Found. Trends Inf. Retr. 2 (12) (2008) 1–135.
[9] S. Poria, E. Cambria, A. Gelbukh, Aspect extraction for opinion mining with a deep convolutional neural network, Knowl.-Based Syst. 108 (2016) 42–49.
[10] Z. Xiaomei, Y. Jing, Z. Jianpei, H. Hongyu, Microblog sentiment analysis with weak dependency connections, Knowl.-Based Syst. 142 (2018) 170–180.
[11] R.K. Amplayo, S. Lee, M. Song, Incorporating product description to sentiment topic models for improved aspect-based sentiment analysis, Inform. Sci. 454–455 (2018) 200–215.
[12] H.H. Do, P. Prasad, A. Maag, A. Alsadoon, Deep learning for aspect-based sentiment analysis: A comparative review, Expert Syst. Appl. 118 (2019) 272–299.
[13] M. Dragoni, M. Federici, A. Rexha, An unsupervised aspect extraction strategy for monitoring real-time reviews stream, Inf. Process. Manage. (2018).
[14] I. Pavlopoulos, Aspect Based Sentiment Analysis Department of Informatics (Ph.D. thesis), Atens University of Economics & Business, 2014.
[15] H. Peng, Y. Ma, Y. Li, E. Cambria, Learning multi-grained aspect target sequence for Chinese sentiment analysis, Knowl.-Based Syst. 148 (2018) 167–176.
[16] M. Pontiki, D. Galanis, H. Papageorgiou, S. Manandhar, I. Androutsopoulos, SemEval-2015 Task 12: Aspect based sentiment analysis, in: Proceedings of the 9th International Workshop on Semantic Evaluation, 2015.
[17] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, SemEval-2014 Task 4: Aspect based sentiment analysis, in: Proceedings of the 8th International Workshop on Semantic Evaluation, 2014, pp. 27–35.
[18] C. Quan, F. Ren, Unsupervised product feature extraction for feature-oriented opinion determination, Inform. Sci. 272 (2014) 16–28.
[19] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, in: Proceedings of the 3rd International Conference on Learning Representations, 2015.

[20] M.T. Luong, H. Pham, C.D. Manning, Effective approaches to attention-based neural machine translation, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 1412–1421.
[21] A. Kumar, O. Irsoy, P. Ondruska, M. Iyyer, J. Bradbury, I. Gulrajani, V. Zhong, R. Paulus, R. Socher, Ask me anything: Dynamic memory networks for natural language processing, in: Proceedings of the 33rd International Conference on Machine Learning, vol. 48, 2016, pp. 1378–1387.
[22] S. Sukhbaatar, A. Szlam, J. Weston, R. Fergus, End-to-end memory networks, Adv. Neural Inf. Process. Syst. 28 (2015) 2440–2448.
[23] A.M. Rush, S. Chopra, J. Weston, A neural attention model for sentence summarization, in: Proceedings of the 2015 Conference on EMNLP, 2015, pp. 379–389.
[24] Y. Bengio, R. Ducharme, P. Vincent, C. Jauvin, A neural probabilistic language model, J. Mach. Learn. Res. (2003) 1137–1155.
[25] L. Jiang, M. Yu, M. Zhou, X. Liu, T. Zhao, Target-dependent twitter sentiment classification, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human LanguageTechnologies, 2011, pp. 151–160.
[26] V. Perez-Rosas, C. Banea, R. Mihalcea, Learning sentiment lexicons in Spanish, LREC (2012) 3077–3081.
[27] M. Hu, B. Liu, Mining and summarizing customer reviews, in: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge discovery and data mining, 2004, pp. 168–177.
[28] Y. Ma, H. Peng, E. Cambria, Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM, in: The Thirty-Second AAAI Conference on Artificial Intelligence, AAAI-18, 2018, pp. 5876–5883.
[29] S. Ruder, P. Ghaffari, J.G. Breslin, INSIGHT-1 at SemEval-2016 Task 5: Deep learning for multilingual aspect-based sentiment analysis, in: Proceedings of the 10th International Workshop on Semantic Evaluation, 2016.
[30] M. Tubishat, N. Idris, M.A.M. Abushariah, Implicit aspect extraction in sentiment analysis: Review, taxonomy, opportunities, and open challenges, Inf. Process. Manag. 54 (4) (2018) 545–563.
[31] Y. Wang, M. Huang, L. Zhao, X. Zhu, Attention-based LSTM for aspect-level sentiment classification, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2016, pp. 606–615.
[32] S. Jebbara, P. Cimiano, Aspect-based sentiment analysis using a two-step neural network architecture, 2017, arXiv:1709.06311, [cs.CL].
[33] X. Li, L. Bing, P. Li, W. Lam, Z. Yang, Aspect term extraction with history attention and selective transformation, in: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18, 2018, pp. 4194–4200.
[34] P. Chen, Z. Sun, L. Bing, W. Yang, Recurrent attention network on memory for aspect sentiment analysis, in: Proceedings of Empirical Methods on Natural Language Processing, 2017, pp. 463–472.
[35] Q. Liu, H. Zhang, Y. Zeng, Z. Huang, Z. Wu, Content attention model for aspect based sentiment analysis, in: Proceedings of the 2018 World Wide Web Conference, 2018, pp. 1023–1032.
[36] D. Tang, B. Qin, T. Liu, Aspect level sentiment classification with deep memory network, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 214–224.
[37] D. Tang, B. Qin, X. Feng, T. Liu, Effective LSTMs for target-dependent sentiment classification, in: International Conference on Computational Linguistics, 2016, pp. 3298–3307.
[38] Y. Tay, L.A. Tuan, S.C. Hui, Dyadic memory networks for aspect-based sentiment analysis, in: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, 2017, pp. 107–116.
[39] A.-M. Popescu, O. Etzioni, Extracting product features and opinions from reviews, in: Proceedings of the Empirical Methods in Natural Language Processing, 2005, pp. 339–346.
[40] M. Van de Kauter, D. Breesch, V. Hoste, Fine-grained analysis of explicit and implicit sentiment in financial news articles, Expert Syst. Appl. 42 (11) (2015) 4999–5010.
[41] S. Jebbara, P. Cimiano, Aspect-based relational sentiment analysis using a stacked neural network architecture, in: Proceedings of the European Conference on Artificial Intelligence, 2016, 1123–1131.
[42] S. Gu, L. Zhang, Y. Hou, Y. Song, A position-aware bi-directional attention network for aspect-level sentiment analysis, in: Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 774–784.
[43] J. Liu, Y. Zhang, Attention modeling for targeted sentiment, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, vol. 2, 2017, pp. 572–577.
[44] D. Ma, S. Li, X. Zhang, H. Wang, Interactive attention networks for aspect-level sentiment classification, in: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, 2017, pp. 4068–4074.
[45] L. Dong, F. Wei, C. Tan, D. Tang, M. Zhou, K. Xu, Adaptive recursive neural network for target-dependent twitter sentiment classification, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014, pp. 49–54.
[46] K. Schouten, F. Frasincar, Survey on aspect-level sentiment analysis, IEEE Trans. Knowl. Data Eng. 28 (3) (2016) 813–830.
[47] Z. Hai, K. Chang, J.-j. Kim, Implicit feature identification via co-occurrence association rule mining, in: Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing, vol. 6608, CICLing 2011, Springer, 2011, pp. 393–404.
[48] C. Long, J. Zhang, X. Zhut, A review selection approach for accurate feature rating estimation, in: Proceedings of the 23rd International Conference on Computational Linguistics, COLING 2010, ACL, 2010, pp. 766–774.
[49] L. Zhang, B. Liu, S.H. Lim, E. O'Brien-Strain, Extracting and ranking product features in opinion documents, in: Proceedings of the 23rd International Conference on Computational Linguistics, COLING 2010, ACL, 2010, pp. 1462–1470.
[50] Y. Zhao, B. Qin, S. Hu, T. Liu, Generalizing syntactic structures for product attribute Candidate extraction, in: Proceedings of the Conference of the North American Chapter of the Association of Computational Linguistics: Human Language Technologies 2010, HLT-NAACL 2010, ACL, 2010, pp. 377–380.
[51] N. Jakob, I. Gurevych, Extracting opinion targets in a single- and cross-domain setting with conditional random fields, in: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP 2010, ACL, 2010, pp. 1035–1045.
[52] Z. Toh, J. Su, NLANGP: Supervised machine learning system for aspect category classification and opinion target extraction, in: Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval 2015, 2015, pp. 496–501.
[53] A.S. Manek, P.D. Shenoy, M.C. Mohan, K.R. Venugopal, Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier, 20 (2) (2017) 135–154.
[54] V. Parkhe, B. B. Biswas, Sentiment analysis of movie reviews: Finding most important movie aspects using driving factors, Soft Comput. 20 (9) (2016) 3373–3379.
[55] A. García-Pablos, M. Cuadros, G. Rigau, W2VLDA: Almost unsupervised system for aspect based sentiment analysis, Expert Syst. Appl. 91 (2018) 127–137.
[56] S. Moghaddam, M. Ester, The FLDA model for aspectbased opinion mining: Addressing the cold start problem, in: Proceedings of the 22nd International Conference on World Wide Web, WWW 2013, ACM, 2013, pp. 909–918.
[57] R.Y.K. Lau, C.C.L. Lai, J. Ma, Y. Li, Automatic domain ontology extraction for context-sensitive opinion mining, in: ICIS Proceedings, vol. 35, 2009.
[58] W. Wei, Jon Atle Gulla, Sentiment learning on product reviews via sentiment ontology tree, in: ACL '10 Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics 2010, pp. 404–413.
[59] J. Yu, Z.-J. Zha, M. Wang, T.-S. Chua, Aspect ranking: Identifying important product aspects from online consumer reviews, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, ACL 2011, ACL, 2011, pp. 1496–1505.
[60] S. Moghaddam, M. Ester, Opinion digger: An unsupervised opinion miner from unstructured product reviews, in: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, CIKM 2010, ACM, 2010, pp. 1825–1828.
[61] Y. Choi, C. Cardie, Learning with compositional semantics as structural inference for subsentential sentiment analysis, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, 2008, pp. 793–801.
[62] M.S. Mubarok, Adiwijaya, M.D. Aldhi, Aspect-based sentiment analysis to review products using Naïve Bayes, AIP Conf. Proc. 1867 (2017) 020060, 1–8.
[63] X. Dong, G. de Melo, A helping hand: Transfer learning for deep sentiment analysis, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018, pp. 2524–2534.
[64] G. Glavaš, I. Vulić, Explicit retrofitting of distributional word vectors, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, vol. 1, 2018, pp. 34–45.
[65] Q. Liu, H. Huang, G. Zhang, Y. Gao, J. Xuan, J. Lu, Semantic structure-based word embedding by incorporating concept convergence and word divergence, in: Proceedings of the Association for the Advancement of Artificial Intelligence, 2018, pp. 5261–5268.
[66] S. Rothe, S. Ebert, H. Schütze, Ultradense word embeddings by orthogonal transformation, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 767–777.
[67] M. Song, H. Park, K. Shin, Attention-based long short-term memory network using sentiment lexicon embedding for aspect-level sentiment analysis in Korean, Inf. Process. Manage. 56 (3) (2019) 637–653.
[68] Q. Wang, Z. Mao, B. Wang, L. Guo, Knowledge graph embedding: A survey of approaches and applications, IEEE Trans. Knowl. Data Eng. 29 (12) (2017) 2724–2743.

[69] S. Xiong, H. Lv, W. Zhao, D. Ji, Towards Twitter sentiment classification by multi-level sentiment-enriched word embedding, Neurocomputing 275 (31) (2018) 2459–2466.

[70] Z. Ye, F. Li, T. Baldwin, Encoding sentiment information into word vectors for sentiment analysis, in: Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 997–1007.

[71] K.M. Hermann, T. Kočiský, E. Grefen-štette, L. Espeholt, W. Kay, M. Suleyman, P. Blunsom, Teaching machines to read and comprehend, 2015, arXiv preprint arXiv:1506.03340.

[72] K. Xu, J. Ba, R. Kiros, A. Courville, R. Salakhutdinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image Caption generation with visual attention, 2015, arXiv preprint arXiv:1502.03044.

[73] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.

[74] T. Young, D. Hazarika, S. Poria, E. Cambria, Recent trends in deep learning based natural language processing, 2017, arXiv:1708.02709.

[75] A. Karpathy, J. Johnson, F. Li, Visualizing and understanding recurrent networks, 2015, arXiv:1506.02078 [cs.LG].

[76] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, in: Proceedings of the Empirical Methods in Natural Language Processing, vol. 14, 2014, pp. 1532–1543.

[77] S. Kiritchenko, X. Zhu, C. Cherry, S.M. Mohammad, NRC-Canada-2014: Detecting aspects and sentiment in customer reviews, in: Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval 2014, 2014, pp. 437–442.

[78] A. Giachanou, F. Crestani, Like it or not: A survey of twitter sentiment analysis methods, ACM Comput. Surv. 49 (2) (2016) 1–41.