Gradual Machine Learning for Aspect-level Sentiment Analysis

Yanyan Wang, Qun Chen, Jiquan Shen, Boyi Hou, Murtadha H. M. Ahmed, Zhanhuai Li

School of Computer Science, Northwestern Polytechnical University

Key Laboratory of Big Data Storage and Management, Northwestern Polytechnical University, Ministry of Industry and Information Technology, Xian Shaanxi,P.R.China

Abstract

The state-of-the-art solutions for Aspect-Level Sentiment Analysis (ALSA) are built on a variety of deep neural networks (DNN), whose efficacy depends on large amounts of accurately labeled training data. Unfortunately, high-quality labeled training data usually require expensive manual work, and are thus not readily available in many real scenarios. In this paper, we aim to enable effective machine labeling for ALSA without the requirement for manual labeling effort. Towards this aim, we present a novel solution based on the recently proposed paradigm of gradual machine learning. It begins with some easy instances in an ALSA task, which can be automatically labeled by the machine with high accuracy, and then gradually labels the more challenging instances by iterative factor graph inference. In the process of gradual machine learning, the hard instances are gradually labeled in small stages based on the estimated evidential certainty provided by the labeled easier instances. Our extensive experiments on the benchmark datasets have shown that the performance of the proposed approach is considerably better than its unsupervised alternatives, and also highly competitive compared to the state-of-the-art supervised DNN techniques.

Keywords: Gradual machine learning, Factor graph inference, Aspect-level sentiment analysis

1. Introduction

Aspect-level sentiment analysis [1], a fine-grained classification task, is highly valuable to both consumers and companies because it can provide the detailed opinions expressed towards certain aspects of an entity. In the literature [2], the task of aspect-level sentiment analysis (ALSA) has been classified into two finer subtasks, aspect-term sentiment analysis (ATSA) and aspect-category sentiment analysis (ACSA). ATSA aims to predict the sentiment polarity associated with the explicit aspect terms appearing in the text. ACSA instead deals with both explicit and implicit aspects. In ACSA, an aspect term may not explicitly appear in the text. However, it needs to predict the sentiment polarities of all the pre-specified aspects in a review. For instance, consider the running example shown in Table 1, in which R_i and S_{ij} denote the review and sentence identifiers respectively. The review R_2 expresses the opinions about a laptop from two aspects, *battery* and *performance*. The goal of ATSA is to predict the sentiment polarity toward the aspect *battery*; while the goal of ACSA is to identify the polarity of both aspects even though the aspect term of *performance* does not appear in the text. In this paper, we target both ATSA and ACSA.

The state-of-the-art techniques for aspect-level sentiment analysis were mainly built on a variety of DNN models [2, 3, 4]. Compared with previous learning models [5, 6], they can effectively improve classification accuracy by automatically learning multiple levels of representations from data. However, their efficacy depends heavily on large amounts of accurately labeled training data. Unfortunately, high-quality labeled data usually require expensive manual work, and are thus not readily available in many real scenarios. To address the limitation resulting from such dependence, this paper presents a novel solution based on the recently proposed paradigm of gradual machine

Email addresses: wangyanyan@mail.nwpu.edu.cn (Yanyan Wang), chenbenben@nwpu.edu.cn (Qun Chen), shenjiquan@mail.nwpu.edu.cn (Jiquan Shen), ntoskrnl@mail.nwpu.edu.cn (Boyi Hou), murtadha@mail.nwpu.edu.cn (Murtadha H. M. Ahmed), lizhh@nwpu.edu.cn (Zhanhuai Li)

Table 1: A Running Example from Laptop Reviews

R_i	S_{ij}	**Text**
R_1	S_{11}	I **like** the battery that can last **long** time.
	S_{12}	However, the keyboard sits a little **far** back for me.
R_2	S_{21}	The laptop has a **long** battery life.
	S_{22}	It also can run my games **smoothly**.

learning [7]. Inspired by the gradual nature of human learning, which is adept at solving problems with increasing hardness, gradual machine learning begins with some easy instances in a task, which can be automatically labeled by the machine with high accuracy, and then gradually reasons about the labels of the more challenging instances based on the observations provided by the labeled instances.

As pointed out in [7, 8], even though there already exist many learning paradigms for a variety of classification tasks, including transfer learning [9], lifelong learning [10], curriculum learning [11], and self-training learning [12] to name a few, the following two properties of gradual machine learning make it fundamentally different from the existing learning paradigms:

- Distribution misalignment between easy and hard instances in a task. The scenario of gradual machine learning does not satisfy the i.i.d (independent and identically distributed) assumption underlying most existing machine learning models: the labeled easy instances are not representative of the unlabeled hard instances. The distribution misalignment between the labeled and unlabeled instances renders most existing learning models unfit for gradual machine learning.

- Gradual learning by small stages in a task. Gradual machine learning proceeds in small stages. At each stage, it typically labels only one instance based on the evidential certainty provided by the labeled easier instances. The process of iterative labeling can be performed in an unsupervised manner without requiring any human intervention.

The major contributions of this paper can be summarized as follows:

- We propose a novel solution for ALSA based on the paradigm of gradual machine learning, which can enable effective machine labeling without the requirement for manual labeling effort. We have presented the detailed techniques for the three steps laid out in the paradigm, which include easy instance labeling, common feature extraction and influence modeling, and scalable gradual inference.

- We have empirically evaluated the performance of the proposed solution by a comparative study. Our evaluation results have shown that its performance is considerably better than the unsupervised alternatives, and also highly competitive compared to the state-of-the-art supervised DNN techniques.

The rest of this paper is organized as follows: Section 2 reviews related work. Section 3 defines the task of ALSA and provides a paradigm overview of gradual machine learning. Section 4 proposes the technical solution for ALSA. Section 5 presents the scalable solution of gradual inference for ALSA. Section 6 empirically evaluates the efficacy of the proposed approach. Finally, we conclude this paper with Section 7.

2. Related work

Machine Learning Paradigms. There exist many machine learning paradigms proposed for a wide variety of classification tasks. Here, our intention is not to exhaustively review all the work. We will instead briefly review those closely related to the paradigm of gradual machine learning used in this paper and discuss their difference.

Traditional machine learning algorithms make predictions on the future data using the statistical models that are trained on previously collected labeled or unlabeled training data [13, 14, 15, 16]. In many real scenarios, the labeled data may be too few to build a good classifier. Semi-supervised learning [17, 18, 19] addresses this problem by making use of a large amount of unlabeled data and a small amount of labeled data. Nevertheless, the efficacy of both supervised and semi-supervised learning paradigms depends on the i.i.d assumption. Therefore, they can not be applied to the scenario of gradual machine learning.

In contrast, transfer learning [9], allows the distributions of the data used in training and testing to be different. It focused on using the labeled training data in a domain to help learning in another target domain. The other learning techniques closely related to transfer learning include lifelong learning [10] and multi-task learning [20]. Lifelong learning is similar to transfer learning in that it also focused on leveraging the experience gained on the past tasks for the current task. However, different from transfer learning, it usually assumes that the current task has good training data, and aims to further improve the learning using both the target domain training data and the knowledge gained in past learning. Multi-task learning instead tries to learn multiple tasks simultaneously even when they are different. A typical approach for multi-task learning is to uncover the pivot features shared among multiple tasks. However, all these learning paradigms can not be applied for the scenario of gradual machine learning. Firstly, focusing on unsupervised learning within a task, gradual machine learning does not enjoy the access to good labeled training data or a well-trained classifier to kick-start learning. Secondly, the existing techniques transfer instances or knowledge between tasks in a batch manner. As a result, they do not support gradual learning by small stages on the instances with increasing hardness within the same task.

The other closely related machine learning paradigms include curriculum learning (CL) [11] and self-paced learning (SPL) [21]. Both of them are similar to gradual machine learning in that they were also inspired by the learning principle underlying the cognitive process in humans, which generally start with learning easier aspects of a task, and then gradually takes more complex examples into consideration. However, both of them depend on a curriculum, which is a sequence of training samples essentially corresponding to a list of samples ranked in ascending order of learning difficulty. A major disparity between them lies in the derivation of the curriculum. In CL, the curriculum is assumed to be given by an oracle beforehand, and remains fixed thereafter. In SPL, the curriculum is instead dynamically generated by the learner itself, according to what the learner has already learned. Based on the traditional learning models, both CL and SPL depend on the i.i.d assumption and require good-coverage training examples for their efficacy. In contrast, the scenario of gradual machine learning does not satisfy the i.i.d assumption. It instead aims to eliminate the dependency on good-coverage training data.

Online learning [22] and incremental learning [23] have also been proposed for the scenarios where training data only becomes available gradually over time or its size is out of system memory limit. Based on the traditional learning models, both online learning and incremental learning depend on high-quality training data for their efficacy. Therefore, they can not be applied for gradual learning either.

Work on Aspect-level Sentiment Analysis. Sentiment analysis at different granularity levels, including document, sentence, and aspect levels, has been extensively studied in the literature [24]. The early work for aspect-level sentiment analysis [6, 5] focused on traditional machine learning algorithms. Unfortunately, the performance of these traditional techniques depends heavily on the quality of the features, which usually have to be manually crafted.

Since deep neural networks can automatically learn high-quality features or representations, most recent work attempted to adapt such models for aspect-level sentiment analysis. The existing work can be divided into two categories based on the two finer subtasks of ATSA and ACSA. For ATSA task, Dong [25] initially proposed an Adaptive Recursive Neural Network (AdaRNN) that can employ a novel multi-compositionality layer to propagate the sentiments of words towards the target. Noticing that the models based on recursive neural network heavily rely on external syntactic parser, which may result in inferior performance, many works subsequently focused on recurrent neural networks. Tang [26] developed a target-dependent LSTM (TD-LSTM) model to capture the connection between target words and their contexts. The alternative solutions include memory networks and convolutional neural networks. Wang [27] proposed Target-sensitive Memory Networks that can capture the sentiment interaction between targets and contexts. Li [28] presented Transformation Networks that employ a CNN layer to extract salient features from the transformed word representations originated from a bi-directional RNN layer. Due to the great success of attention mechanism in image recognition [29], speech recognition [30], machine translation [31, 32] and question answering [33], many models based on LSTM and attention mechanism have also been proposed. These models, including Hierarchical Attention Network [34], Segmentation Attention Network [35], Interactive Attention Networks [36], Recurrent Attention Network [37], Attention-over-Attention Neural Networks [38], Effective Attention Modeling [39], Content Attention Model [40], Multi-grained Attention Network [41], employed different attention mechanisms to output the aspect-specific sentiment features.

Table 2: Frequently used notations.

Notation	Description
$t_i = (r_j, s_k, a_l)$	an aspect unit
r_j	a review
s_k	a sentence
a_l	an aspect category or aspect term
$T = \{t_i\}$	a set of aspect units

In comparison, there exist fewer works for ACSA because the implicit aspects make the task more challenging. Ruder [4] proposed a hierarchical bidirectional LSTM for the ACSA task by modeling the inter-dependencies of sentences in a review, which does not fully employ the given aspect. Wang [3] presented an attention-based LSTM that employs an aspect-to-sentence attention mechanism to concentrate on the key part of a sentence given an aspect. Xue [2] introduced a model based on convolutional neural networks and gating mechanisms, which is more accurate and efficient. Note that the models proposed for ACSA can also be used for ATSA, but the ones for ATSA usually solely benefit themselves because they usually employ specific components to model explicit aspect-term together with its relative context.

Unfortunately, the efficacy of the aforementioned DNN-based approaches depends heavily on good-coverage training data, which may not be readily available in many real scenarios.

3. Preliminaries

In this section, we first define the task of aspect-level sentiment analysis, and then provide a paradigm overview of gradual machine learning.

3.1. Task Statement

For presentation simplicity, we have summarized the frequently used notations in Table 2. Formally, we formulate the task of aspect-level sentiment analysis as follows:

Definition 1. [Aspect-level Sentiment Analysis] *Let $t_i = (r_j, s_k, a_l)$ be an aspect unit, where r_j is a review, s_k is a sentence in the review r_j, and a_l is an aspect associated with the sentence s_k. Note that the aspect a_l can be a aspect category or aspect term, and a sentence may express opinions towards multiple aspects. Given a corpus of reviews, R, the goal of the task is to predict the sentiment polarity of each aspect unit t_i in R.*

In this paper, we suppose that an aspect polarity is either *positive* or *negative*.

3.2. Paradigm overview

The paradigm of gradual machine learning, as shown in Figure 1, consists of the following three steps:

- **Easy Instance Labeling.** Given a classification task, it is usually very challenging to accurately label all the instances in the task without good-coverage training examples. However, the work can become much easier if we only need to automatically label some easy instances in the task. In real scenarios, easy instance labeling can be performed based on the simple user-specified rules or the existing unsupervised learning techniques. For instance, in unsupervised clustering, an instance close to the center of a cluster in the feature space can be considered as an easy instance, because it has only a remote chance to be misclassified. Gradual machine learning begins with the observations provided by the labels of easy instances. Therefore, high accuracy of automatic machine labeling on easy instances is critical for its ultimate performance on a given task.
- **Feature Extraction and Influence Modeling.** Features serve as the medium to convey the knowledge obtained from the labeled easy instances to the unlabeled harder ones. This step extracts the common features shared by the labeled and unlabeled instances. To facilitate effective knowledge conveyance, it is desirable that a wide variety of features are extracted to capture as much information as possible. For each extracted feature, this step also needs to model its influence over the labels of its relevant instances.

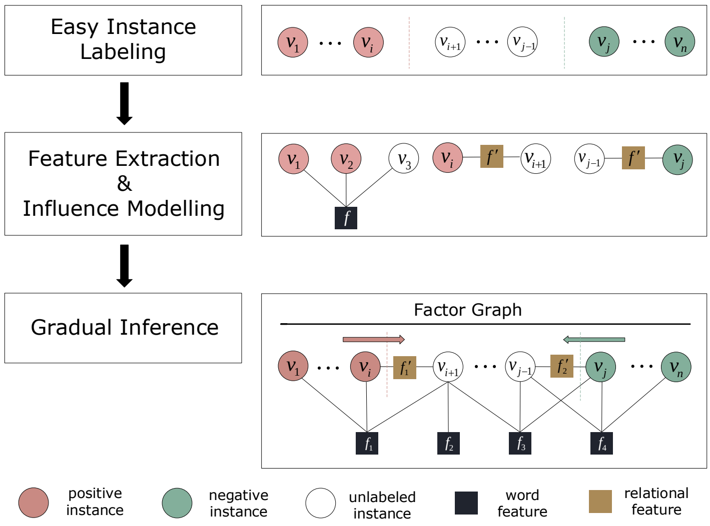

Figure 1: Learning Paradigm Overview.

- **Gradual Inference.** This step gradually labels the instances with increasing hardness in a task. Since the scenario of gradual learning does not satisfy the i.i.d assumption, gradual learning is fulfilled from the perspective of evidential certainty. As shown in Figure 1, it constructs a factor graph, which consists of the labeled and unlabeled instances and their common features. Gradual learning is conducted over the factor graph by iterative inference. At each iteration, it chooses to label the unlabeled instance with the highest degree of evidential certainty. The iteration is repeatedly invoked until all the instances in a task are labeled. In gradual inference, a newly labeled instance at the current iteration would serve as an evidence observation in the following iterations.

4. Solution for ALSA

In this section, we present the solution of gradual machine learning for ALSA. We will present the respective techniques for the three steps laid out in Subsection 3.2.

4.1. Easy Instance Labeling

The existing lexicon-based approaches essentially reason about polarity by summing up the polarity scores of all the sentiment words in a sentence. The score of a sentiment word indicates its intensity of sentiment, which increases with the absolute value of score. Since negation words can effectively reverse the polarity of a sentiment word, they usually perform negation detection for each sentiment word by examining whether there is any negation in its neighboring words [42].

Unfortunately, the existing lexicon-based approaches [43] are prone to error under some ambiguous circumstances. Firstly, the presence of contrast (e.g. *but* and *although*), hypothetical (e.g. *if*) or condition (e.g. *unless*) connectives could significantly complicate polarity detection. Secondly, the presence of negation words involving long-distance dependency could also make the detection challenging. Finally, a sentence may contain multiple sentiment words that hold conflicting polarities; in this case, the true polarity of a sentence is not easily detectable based on sentiment word scoring.

Therefore, we define an easy instance by excluding the aforementioned ambiguous circumstances as follows:

Definition 2. [Easy Instance] *We consider an aspect polarity as an easy instance if and only if the sentence expressing opinions about the aspect simultaneously satisfies the following three conditions:*

- *It contains at least one sentiment word, but does not simultaneously contain any sentiment word holding a conflicting polarity;*

- *It does not contain any contrast, hypothetical or condition connective;*
- *It does not contain any negation word involving long-distance dependency.*

The polarity of an easy instance is simply determined by the polarity of its sentiment words. Moreover, a negation word is supposed to involve long-distance dependency if and only if it is not in the neighboring 3-grams preceding any sentiment word. We illustrate the difference between the easy and challenging instances by Example 1.

Example 1. [Easy Instances] *In a phone review, the sentence "the screen is not good for carrying around in your bare hands ", expressing opinion about "screen", is an easy instance because the sentiment word "good" associated with the local negation cue "not" strongly indicates the negative sentiment. In contrast, the sentence "I don't know why anyone would want to write a great review about this battery", expressing opinions about "battery", is not an easy instance. Even though it contains the strong sentiment word "great", it also includes the negation word "don't" involving long-distance dependency. Similarly, the sentence "I like this laptop, the only problem is that it can not last long time" is not an easy instance, because it contains both the positive and negative words (e.g. "like" and "problem").*

4.2. Feature Extraction and Influence Modeling

We extract two types of features for influence modeling: word feature and relational feature.

Word Feature. We extract sentiment words, which are specified in open-source lexicons, from sentences and consider them as features of aspect polarities. To capture more information shared among aspect instances, we also extract k-grams ($k \geq 2$) as the word features besides the single sentiment word. Since negation words can effectively reverse the polarity of a sentiment word [44, 45], we perform negation detection for each sentiment word by examining whether there is any negation in its neighboring words. Note that a sentence may express opinions towards multiple aspects. In this case, firstly, we employ the dependency-based parse tree [46] to extract all the pairs of opinion target and opinion word [47], and assign an opinion word to an aspect if either itself or its opinion target is close to the aspect in the vector space (namely their similarity is less than a threshold (e.g. 0.5)). Secondly, we extract the opinion phrases towards an aspect based on its opinion words, and assign the k-gram features contained in the opinion phrases to the aspect.

Relational Feature. The sentences in a review build upon each other. There often exist some discourse relations between clauses or sentences, which can provide valuable hints for polarity reasoning. Specifically, it can be observed that two sentences connected with a shift word usually have opposite polarities. In contrary, two neighboring sentences without any shift word between them usually have similar polarities. In the running example shown in Table 1, the polarities of S_{11} and S_{12} are opposite because they are connected by the shift word of "but", while the polarities of S_{21} and S_{22} are similar due to the absence of any shift word between them.

Therefore, we employ the rules to extract the similar and opposite relations between aspect polarity based on sentence context. Given two aspect units $t_i = \{r_i, s_i, a_i\}$ and $t_j = \{r_j, s_j, a_j\}$ that are opinioned in the same review (namely $r_i = r_j$), the rules for polarity relations extraction are as follows:

- If the sentences s_i and s_j are identical (s_i=s_j) or adjacent and neither of them contains any shift word, t_i and t_j are supposed to hold similar polarities;
- If two adjacent sentences s_i and s_j are connected by a shift word and neither of them contains any inner-sentence shift word, t_i and t_j are supposed to hold opposite polarities;
- If the sentences s_i and s_j are identical and the opinion clauses associated with them are connected by a inner-sentence shift word, t_i and t_j are supposed to hold opposite polarities.

Note that given an ATSA task, it is easy to identify the opinion clause because the aspect term explicitly appears in the text. Therefore, the 3rd rule can be easily checked in the scenario of ATSA. The scenario of ACSA is instead more challenging. Our solution first uses the dependency-based parse tree to extract all the pairs of opinion phrases, and associates an opinion clause with a specific aspect if either its opinion target or opinion word is close to the aspect in the vector space.

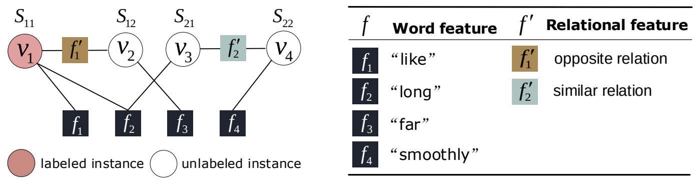

Figure 2: Factor Graph for the Running Example.

4.3. Gradual Inference

To enable gradual learning, we construct a factor graph, G, in which the labeled easy instances are represented by the *evidence variables*, the unlabeled hard instances by the *inference variables*, and the features by the *factors*. The value of each variable represents its corresponding polarity. An evidence variable has the constant value of 0 or 1, which indicate the polarity of *negative* and *positive* respectively. The values of evidence variables remain unchanged during the inference process. The values of the inference variables should instead be inferred based on G. The factor graph constructed for the running example is shown in Figure 2.

Gradual machine learning is attained by iterative factor graph inference on G. In G, we define the probability distribution over its variables V by

$$P_w(V) = \frac{1}{Z_w} \prod_{v \in V} \prod_{f \in F_v} \phi_f(v) \prod_{f' \in F'} \phi_{f'}(v_i, v_j), \tag{1}$$

where F_v denotes the set of word features associated with the variable v, F' denotes the set of relational features, $\phi_f(v)$ denotes the factor associated with v and f, and $\phi_{f'}(v_i, v_j)$ denotes the factor associated with the relational feature f'. In Eq. 1, the factor of a word feature f is defined by

$$\phi_f(v_i) = \begin{cases} 1 & v_i = 0; \\ e^{w_f} & v_i = 1; \end{cases} \tag{2}$$

where v_i denotes a variable having the feature f, and w_f denotes the weight of f. Similarly, the factor of a relational feature f' is defined by

$$\phi_{f'}(v_i, v_j) = \begin{cases} e^{w_{f'}} & if\ v_i = v_j; \\ 1 & otherwise; \end{cases} \tag{3}$$

where v_i and v_j denote the two variables sharing the feature f', and $w_{f'}$ denotes the weight of f'.

As in [7, 8], given a factor graph with some labeled evidence variables, we reason about the factor weights by minimizing the negative log marginal likelihood of

$$\hat{w} = arg \min_{w} -log \sum_{V_I} P_w(\Lambda, V_I), \tag{4}$$

where Λ denotes the observed labels of evidence variables and V_I denotes the set of inference variables. The objective function effectively learns the factor weights most consistent with the label observations of the evidence variables.

As usual, gradual inference proceeds in small stages. At each stage, it chooses to label the unlabeled variable with the highest degree of evidential certainty in G. The iteration is repeatedly invoked until all the inference variables are labeled. In gradual inference, evidential certainty is measured by the inverse of entropy. Formally, entropy is formally defined by

$$H(v) = -(P(v) \cdot \ln P(v) + (1 - P(v)) \cdot \ln(1 - P(v))), \tag{5}$$

in which $H(v)$ denotes the entropy of a variable v.

In our implementation, we use the Numbskull library [1] to optimize this objective function by interleaving stochastic gradient descent steps with Gibbs sampling ones, similar to contrastive divergence. However, repeated inference by maximum likelihood over a large-sized factor graph is usually time-consuming [48]. Therefore, in the next section, we will present a scalable solution for gradual inference on ALSA.

[1] https://github.com/HazyResearch/numbskull

Algorithm 1: Scalable Gradual Inference

while *there exists any unlabeled variable in* G **do**
    $V' \leftarrow$ all the unlabeled variables in G;
    for $v \in V'$ **do**
        Measure the evidential support of v in G;
    end
    Select top-m unlabeled variables with the most evidential support (denoted by V_m) ;
    for $v \in V_m$ **do**
        Approximately rank the entropy of v in V_m;
    end
    Select top-k most promising variables in terms of entropy in V_m (denoted by V_k) ;
    for $v \in V_k$ **do**
        Compute the probability of v in G by factor graph inference over a subgraph of G;
    end
    Label the variable with the minimal entropy in V_k;
end

5. Scalable Gradual Inference

The scalable solution is built based on the framework proposed in [8]. It consists of three steps, measurement of evidential support, approximate ranking of entropy and construction of inference subgraph. Its efficacy depends on the following observations:

- Many unlabeled inference variables in the factor graph are only weakly linked through the factors to the evidence variables. Due to the lack of evidential support, their inferred probabilities would be quite ambiguous, i.e. close to 0.5. As a result, only the inference variables that receive considerable support from the evidence variables need to be considered for labeling;
- With regard to the probability inference of a single variable v in a large factor graph, it can be effectively approximated by considering the potentially much smaller subgraph consisting of v and its neighboring variables. The inference over the subgraph can usually be much more efficient than over the original entire graph.

The process of scalable gradual inference for ALSA is sketched in Algorithm 1. It first selects the top-m unlabeled variables with the most evidential support in G as the candidates for probability inference. To reduce the invocation frequency of factor graph inference, it then approximates entropy estimation by an efficient algorithm on the m candidates and selects only the top-k most promising variables among them for factor graph inference. Finally, it infers the probabilities of the chosen k variables in G. For each variable, its probability is not inferred over the entire graph of G, but over a potentially much smaller subgraph.

5.1. Measurement of Evidential Support

Our solution for measuring evidential support is built based on the Dempster-Shafer (D-S) theory [49]. Given an inference variable, we first estimate the evidential support provided by each of its factors, and then aggregate them to measure its overall evidential support.

First, we describe how to estimate the evidential support provided by a single factor. Formally, let X be the universal set representing all possible states of a system under consideration. By a function of basic belief assignment, the D-S theory assigns a belief mass to each element of the power set of X. The mass of an element E, denoted by $m(E)$, expresses the proportion of all relevant and available evidence that supports the claim that the actual state belongs to E but to no particular subset of E. The masses of elements satisfy

$$\sum_{E \in 2^X} m(E) = 1 \;\; \& \;\; m(\emptyset) = 0. \tag{6}$$

In case that only singleton propositions are assigned belief masses, a mass function reduces to a classical probability function.

For evidential support measurement, we define two propositions: "label the instance", denoted by L, and "unlabel the instance, denoted by U. With $X = \{L, U\}$, the power set of X can be represented by $2^X = \{\emptyset, L, U, X\}$. Given an inference variable v and its word feature f, we estimate the evidential support provided by f by the mass function

$$m_f(E) = \begin{cases} (1 - d_f) \cdot max\{P(f), 1 - P(f)\} & E = \{L\}, \\ (1 - d_f) \cdot min\{P(f), 1 - P(f)\} & E = \{U\}, \\ d_f & E = \{L, U\}, \end{cases} \tag{7}$$

where d_f denotes the degree of uncertainty of f, and $P(f)$ denotes the proportion of positive instances among all labeled instances having the feature f. According to Eq. 7, the mass assigned to the element of $\{L\}$ increases as the value of $P(f)$ becomes more extreme (i.e. close to 0 or 1). The underlying intuition is that the more extreme the value of $P(f)$ is, the more evidential support the element of $\{L\}$ should receive from the feature f.

Similarly, given an inference variable v and its relational feature f', we estimate the evidential support provided by f' by the mass function

$$m_{f'}(E) = \begin{cases} (1 - d_{f'}) \cdot R(f') & E = \{L\}, \\ (1 - d_{f'}) \cdot (1 - R(f')) & E = \{U\}, \\ d_{f'} & E = \{L, U\}, \end{cases} \tag{8}$$

where $d_{f'}$ denotes the degree of uncertainty of f', and $R(f')$ denotes the accuracy of the relation f'. In Eq. 8, $R(f')$ can be considered as the statistical accuracy of the extracted relations, and the evidential support the element of $\{L\}$ receives from f' increases with the estimated accuracy.

Next, we describe how to measure the aggregate evidential support provided by multiple factors. Suppose that an inference variable v has i word features, $\{f_1, \ldots, f_i\}$, and j relational features, $\{f'_1, \ldots, f'_j\}$. Given the element of $E = \{L\}$, we estimate its aggregate evident support by combining the estimated masses as follows

$$m(E) = m_{f_1}(E) \oplus \cdots \oplus m_{f_i}(E) \oplus m_{f'_1}(E) \oplus \cdots \oplus m_{f'_j}(E), \tag{9}$$

where $m(E)$ denotes the total amount of evidential support v receives, and the combination is calculated from the two sets of mass functions, $m_{f_1}(E)$ and $m_{f_2}(E)$, in the following manner

$$m_{f_1}(E) \oplus m_{f_2}(E) = \frac{1}{1 - K} \sum_{E' \cap E'' = E} m_{f_1}(E') \cdot m_{f_2}(E''), \tag{10}$$

where E' and E'' denote the elements of the power set, and

$$K = \sum_{E' \cap E'' = \emptyset} m_{f_1}(E') \cdot m_{f_2}(E''), \tag{11}$$

which is a measure of the amount of conflict between E' and E''.

Note that the degree of uncertainty, denoted by d_f and $d_{f'}$ in Eq. 7 and 8 respectively, indicates how much impact a feature has on the whole degree of belief in terms of evidential support measurement. The lower the value, the greater the impact. It can be observed that relational features can usually provide more reliable information than word features. In practical implementation, we suggest that $d_{f'}$ is set to be smaller than d_f (e.g., $d_f = 0.4$ and $d_{f'} = 0.1$). Our empirical evaluation in Subsection 6.3 has shown that the performance of gradual machine learning is, to a large extent, insensitive to the parameter setting of d_f and $d_{f'}$.

5.2. *Approximate Ranking of Entropy*

Since more evidential conflict means more status uncertainty, we approximate the entropy ranking of inference variables by measuring their evidential conflict. Similar to the case of evidential support measurement, we define two propositions: "label it as *positive*", denoted by L^+, and "label it as *negative*", denoted by L^-. Given an inference variable v and its word feature f, we approximate v's evidential certainty w.r.t f with the mass function

$$m^*_f(E) = \begin{cases} (1 - d^*_f) \cdot P(f) & E = \{L^+\}, \\ (1 - d^*_f) \cdot (1 - P(f)) & E = \{L^-\}, \\ d^*_f & E = \{L^+, L^-\}, \end{cases} \tag{12}$$

where d^*_f denotes the degree of uncertainty of f, and $P(f)$ denotes the proportion of positive instances among all the labeled instances having the feature f.

Similarly, given an inference variable v and its relational feature f', we approximate v's evidential certainty w.r.t f' with the mass function

$$m^*_{f'}(E) = \begin{cases} (1 - d^*_{f'}) \cdot P(f') & E = \{L^+\}, \\ (1 - d^*_{f'}) \cdot (1 - P(f')) & E = \{L^-\}, \\ d^*_{f'} & E = \{L^+, L^-\}, \end{cases} \tag{13}$$

where $d^*_{f'}$ denotes the degree of uncertainty of f', and $P(f')$ denotes the probability of v being positive if only the evidence f' is considered for labeling v. If the labeled variable on the other side of the relation f' is positive, we set $P(f') = \frac{e^{w_{f'}}}{1+e^{w_{f'}}}$, in which $w_{f'}$ denotes the weight of f'; otherwise (it is negative), we set $P(f') = \frac{1}{1+e^{w_{f'}}}$.

Finally, we measure the amount of conflict between the multiple pieces of evidence using the generalized expression of K as specified in Eq. 11. Similar to the case of evidential support measurement, we suggest that $d^*_{f'}$ is set to be a lower value than d^*_f. Our empirical evaluation in Subsection 6.3 has shown that the performance of gradual machine learning is, to a large extent, insensitive to the parameter setting of d^*_f and $d^*_{f'}$.

5.3. Construction of Inference Subgraph

It has been empirically shown [50] that given a variable v in G, its probability inference can be effectively approximated by considering the subgraph only consisting of v and its r-hop neighboring variables, and even with a small value of r (e.g. 2 and 3), the approximation can be sufficiently accurate in many real scenarios. Therefore, given a target inference variable v in G, we first extract all its 2-hop neighbors reachable by the relational factors and include them in the subgraph. For each word feature of v, all the labeled and unlabeled instances sharing the feature with v are also included in the constructed subgraph because their labels may play important roles in labeling v through the common word feature.

6. Empirical Evaluation

In this section, we empirically evaluate the performance of the proposed solution by a comparative study. We compared GML with the state-of-the-art methods proposed for both ACSA and ATSA. For the ACSA task, the alternative techniques include:

- **LEX-SYN [51].** It is an unsupervised approach built on lexicons and syntactic dependency analysis;
- **H-LSTM [4].** It is an enhanced DNN technique. It models the inter-dependencies of sentences in a review using a hierarchical bidirectional LSTM;
- **AT-LSTM [3].** Referring to the Attention-based LSTM, it employs an attention mechanism to concentrate on the key parts of a sentence given an aspect, where the aspect embeddings are used to determine the attention weights;
- **ATAE-LSTM [3].** Referring to the Attention-based LSTM with Aspect Embedding, it is supposed to be an improvement over AT-LSTM. It extends AT-LSTM by appending the input aspect embedding into each word's input vector.
- **GCAE [2].** It is a model based on convolutional neural networks and gating mechanisms that can selectively output the sentiment features accordingto a given aspect.

For the ATSA task, besides LEX-SYN, AT-LSTM, ATAE-LSTM and GCAE, the alternative techniques also include:

- **IAN [36].** Referring to the interactive attention network, it can interactively learn attentions in the contexts and targets and generate the representations for targets and contexts separately;
- **RAM [37].** It is a multiple-attention network that can effectively capture sentiment features separated by a long distance, and is usually more robust against irrelevant information;

Table 3: Accuracy Comparison for ACSA on Benchmark Datasets

Model	PHO16	CAM16	LAP16	RES16	LAP15	RES15
LEX-SYN	67.11%	74.64%	68.18%	73.74%	70.97%	71.07%
H-LSTM	73.30%	78.80%	78.90%	83.10%	80.00%	77.10%
AT-LSTM	72.40%	81.70%	76.03%	85.03%	81.03%	77.25%
ATAE-LSTM	74.48%	83.36%	79.07%	84.66%	80.68%	79.13%
GCAE	**77.69%**	**83.78%**	**80.42%**	**87.45%**	82.26%	**80.28%**
GML	76.16%	81.46%	80.40%	85.37%	**83.82%**	78.78%

- **AOA [38].** Referring to the attention-over-attention network, it models aspects and sentences in a joint way and can explicitly capture the interaction between aspects and context sentences;
- **TNet [28].** It is a target-specific transformation network taht can better integrate target information into the word representations.

The rest of this section is organized as follows: Subsection 6.1 describes the experimental setup; Subsection 6.2 presents the comparative evaluations results; Subsection 6.3 evaluates the performance sensitivity of the proposed solution w.r.t various parameters.

6.1. Experimental Setup

In the empirical evaluation, we have used six benchmark datasets in four domains (phone, camera, laptop and restaurant) and two languages (English and Chinese) from the SemEval 2015 task 12 [52] and 2016 task 5 [53]. In all the experiments, we perform 2-class classification to label an aspect polarity as *positive* or *negative*. Note that the datasets of LAP16, RES16, LAP15 and RES15 contain some neutral instances, which are simply ignored in our experiments. There are no labeled aspect terms in the Chinese datasets of PHO16 and CAM16. Therefore, for ATSA, we only compare GML to its alternatives on the English datasets.

For DNN models, we used Glove embeddings [2] for English data, and word embeddings from Baidu [3] for Chinese data. We employed jieba [4] to tokenize Chinese sentences. In easy instance labeling and feature extraction for GML, we used the open-source Opinion Lexicon [5] for English data, and employ EmotionOntology [6] and BosonNLP [7] lexicons for Chinese data. For easy instance identification, the scores for sentiment words in Chinese lexicon are normalized into the range of [-4, 4], and we use the sentiment words whose scores are at least 1. In the process of scalable gradual inference for ALSA, if none of the unlabeled instances receives any evidential support from the labeled easier instances, GML employs the existing unsupervised method [51] to label its polarity. Note that in our implementations of GML, the initial weights of word features, similar relational features and opposite relational features are set to 0, 2 and -2 respectively.

All the reported results are the averages over ten runs. Our implementation codes have been made open-source available at the website[8].

6.2. Comparative Evaluation

In comparative evaluation, we set m=20, k=3, d_f=d^*_f=0.4, and $d_{f'}$=$d^*_{f'}$=0.1 for GML. The detailed evaluation results for both ACSA and ATSA are presented in Table 3 and 4 respectively. The best result achieved on each dataset is also highlighted in the table.

We can see that GML consistently outperforms the unsupervised alternative, LEX-SYN, by considerable margins on all the benchmark datasets. For ACSA, the improvement margins on *PHO16*, *CAM16* and *RES15* are around 7-9%;

[2] https://nlp.stanford.edu/projects/glove/
[3] http://pan.baidu.com/s/1jIb3yr8
[4] https://github.com/fxsjy/jieba
[5] https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html
[6] http://ir.dlut.edu.cn/EmotionOntologyDownload
[7] https://bosonnlp.com/dev/resource
[8] http://www.wowbigdata.com.cn/GML-ALSA/GML-ALSA.html

Table 4: Accuracy comparison for ATSA on benchmark datasets.

Model	LAP16	RES16	LAP15	RES15
LEX-SYN	68.28%	80.97%	68.78%	65.80%
AT-LSTM	74.85%	84.43%	77.51%	75.43%
ATAE-LSTM	75.08%	84.60%	77.66%	74.13%
GCAE	78.34%	**88.86%**	**81.37%**	77.60%
IAN	74.02%	85.12%	79.27%	75.00%
RAM	77.47%	85.81%	78.58%	73.23%
AOA	74.94%	87.02%	80.73%	73.43%
TNet	75.86%	87.20%	80.00%	75.20%
GML	**78.62%**	85.93%	80.89%	**79.84%**

the margins on *LAP16*, *RES16* and *LAP15* are the largest at more than 10%. For ATSA, it achieves the improvement of more than 10% on three out of totally four datasets (namely *LAP16*, *LAP15* and *RES15*). Due to the widely recognized challenge of sentiment analysis, the achieved improvements can be deemed very considerable.

Furthermore, it can be observed that the performance of GML is highly competitive compared to the supervised DNN techniques. Except GCAE, GML achieves overall better performance than all the other DNN models. For instance, for ACSA, GML beats AT-LSTM in performance on five out of totally six datasets; it also beats ATAE-LSTM in performance on four out of the six datasets, and their performance on *RES15* is very close. For ATSA, GML achieves the best performance on two out of totally four datasets; except GCAE, it outperforms all the other DNN model on at least three out of the four datasets. Even compared to GCAE, GML beats it in performance on *LAP15* for ACSA, and *LAP16* and *RES15* for ATSA, and their performance on the other datasets are close. *It is worthy to point out that unlike the DNN models, GML does not use any labeled training data provided in the benchmark. These experimental results evidently demonstrate the efficacy of GML.*

6.3. Sensitivity Evaluation

Table 5: Sensitivity evaluation over ACSA task.

		PHO16	CAM16	LAP16	RES16	LAP15	RES15
w.r.t $\boldsymbol{m}$ (k = 3)	m = 10	75.91%	81.54%	79.95%	83.93%	84.24%	79.39%
	m = 20	76.16%	81.46%	80.40%	85.37%	83.82%	78.78%
	m = 30	76.26%	81.25%	80.08%	84.79%	83.94%	79.47%
	m = 40	75.61%	80.96%	80.08%	84.76%	84.03%	79.11%
w.r.t $\boldsymbol{k}$ (m = 20)	k = 1	77.28%	81.58%	79.95%	84.99%	83.06%	79.81%
	k = 3	76.16%	81.46%	80.40%	85.37%	83.82%	78.78%
	k = 5	75.69%	80.91%	77.90%	85.11%	82.97%	79.08%
	k = 7	75.92%	80.54%	80.05%	85.08%	83.16%	78.69%

	$d_{f'}$ d_f	PHO16	CAM16	LAP16	RES16	LAP15	RES15
w.r.t $\boldsymbol{d_{f'}}$ and $\boldsymbol{d_f}$ (m = 20 k = 3)	0.1 0.2	76.60%	81.25%	79.81%	82.91%	84.22%	79.64%
	0.1 0.3	76.67%	80.79%	79.76%	84.53%	83.96%	78.64%
	0.1 0.4	76.16%	81.46%	80.40%	85.37%	83.82%	78.78%
	0.2 0.3	76.45%	80.91%	79.73%	84.82%	84.08%	78.97%
	0.2 0.4	76.41%	81.08%	79.89%	84.59%	84.01%	78.72%
	0.2 0.5	76.26%	81.41%	79.81%	85.05%	84.12%	79.69%
	0.3 0.4	75.95%	80.96%	79.79%	84.62%	83.99%	78.92%
	0.3 0.5	76.22%	80.91%	80.16%	85.25%	83.64%	78.58%

In sensitivity evaluation, we first vary the values of the parameters m and k as shown in Algorithm 1, which respectively denote the number of candidate variables selected for approximate entropy ranking and the number of candidate variables selected for factor graph inference. We set m=10, 20, 30, 40, and k=1, 3, 5, 7. We then vary the values of the parameters, d_f, $d_{f'}$, d^*_f and $d^*_{f'}$, which denote the degree of uncertainty of word and relational features. We set d_f=d^*_f, $d_{f'} = d^*_{f'}$, and $d_{f'} < d_f <= 0.5$. While evaluating GML's sensitivity to a particular parameter, any other parameter is fixed to the same value.

The detailed evaluation results on ACSA are presented in Table 5. Note that the evaluation results on ATSA are similar, thus omitted here. It can be observed that the performance of GML only fluctuate slightly ($\leq$ 1% in most cases) with different parameter settings. It is worthy to note that the performance of GML does not tend to change much with various values of m and k. Since most of GML's runtime is spent on factor graph inference, reducing the value of k can effectively improve efficiency. Our experiments show that even with k taking the minimal value of 1, the performance of GML only changes marginally. We also have the similar observation on the parameter setting of d_f and $d_{f'}$. Various value combinations of d_f and $d_{f'}$ can only result in very marginal performance fluctuations. Our experimental results clearly show that the performance of GML is, to a large extent, insensitive to the parameter settings. They bode well for its applicability in real scenarios.

7. Conclusion

In this paper, we have proposed a novel paradigm of gradual machine learning and developed a corresponding solution for the task of ALSA. Our empirical study on the benchmark datasets has validated the efficacy of the proposed approach. Using ALSA as a test case, we have demonstrated that gradual machine learning is a promising paradigm.

Future work can be pursued on several fronts. Since GML does not require labeled training data, additional corpora of reviews can be potentially integrated into the inference process for improved performance. Detailed study on this topic is beyond the scope of this paper, but it deserves an independent investigation. It is also interesting to develop solutions of gradual machine learning for other challenging classification tasks requiring extensive labeling effort.

References

References

[1] K. Schouten, F. Frasincar, Survey on aspect-level sentiment analysis, IEEE Transactions on Knowledge and Data Engineering 28 (3) (2016) 813–830.

[2] W. Xue, T. Li, Aspect based sentiment analysis with gated convolutional networks, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL, Melbourne, Australia, 2018, pp. 2514–2523.

[3] Y. Wang, M. Huang, X. Zhu, L. Zhao, Attention-based LSTM for aspect-level sentiment classification, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP, Austin, Texas, USA, 2016, pp. 606–615.

[4] S. Ruder, P. Ghaffari, J. G. Breslin, A hierarchical model of reviews for aspect-based sentiment analysis, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP, Austin, Texas, USA, 2016, pp. 999–1005.

[5] S. Kiritchenko, X. Zhu, C. Cherry, S. Mohammad, Nrc-canada-2014: detecting aspects and sentiment in customer reviews, in: Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval@COLING, Dublin, Ireland, 2014, pp. 437–442.

[6] J. Saias, Sentiue: target and aspect based sentiment analysis in semeval-2015 task 12, in: Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT, Denver, Colorado, USA, 2015, pp. 767–771.

[7] B. Hou, Q. Chen, J. Shen, X. Liu, P. Zhong, Y. Wang, Z. Chen, Z. Li, Gradual machine learning for entity resolution, in: Proceedings of the World Wide Web Conference (Demonstrations), WWW 2019, San Francisco, CA, USA, 2019, pp. 3526–3530.

[8] B. Hou, Q. Chen, Y. Wang, P. Zhong, M. H. M. Ahmed, Z. Chen, Z. Li, Gradual machine learning for entity resolution, CoRR abs/1810.12125 (under review).

[9] S. J. Pan, Q. Yang, A survey on transfer learning, IEEE Transactions on Knowledge and Data Engineering 22 (10) (2010) 1345–1359.

[10] Z. Chen, B. Liu, Lifelong machine learning, second edition, Synthesis Lectures on Artificial Intelligence and Machine Learning, Morgan & Claypool Publishers, 2018.

[11] Y. Bengio, J. Louradour, R. Collobert, J. Weston, Curriculum learning, in: Proceedings of the 26th Annual International Conference on Machine Learning, ICML, Montreal, Quebec, Canada, 2009, pp. 41–48.

[12] R. Mihalcea, Co-training and self-training for word sense disambiguation, in: Proceedings of the Eighth Conference on Computational Natural Language Learning, CoNLL, Boston, Massachusetts, USA, 2004, pp. 33–40.

[13] P. Christen, Automatic record linkage using seeded nearest neighbour and support vector machine classification, in: Proceedings of the 14th ACM International Conference on Knowledge Discovery and Data Mining, KDD, Las Vegas, Nevada, USA, 2008, pp. 151–159.

[14] T. Mullen, N. Collier, Sentiment analysis using support vector machines with diverse information sources, in: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing , EMNLP, Barcelona, Spain, 2004, pp. 412–418.
[15] A. Blum, T. M. Mitchell, Combining labeled and unlabeled data with co-training, in: Proceedings of the 17th Annual Conference on Computational Learning Theory, COLT, Madison, Wisconsin, USA, 1998, pp. 92–100.
[16] K. Bellare, S. Iyengar, A. G. Parameswaran, V. Rastogi, Active sampling for entity matching, in: Proceedings of 18th ACM International Conference on Knowledge Discovery and Data Mining, KDD, Beijing, China, 2012, pp. 1131–1139.
[17] K. Nigam, A. McCallum, S. Thrun, T. M. Mitchell, Text classification from labeled and unlabeled documents using EM, Machine Learning 39 (2) (2000) 103–134.
[18] B. Xiang, L. Zhou, Improving twitter sentiment analysis with topic-based mixture modeling and semi-supervised training, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL, Baltimore, MD, USA, 2014, pp. 434–439.
[19] H. He, X. Sun, A unified model for cross-domain and semi-supervised named entity recognition in chinese social media, in: Proceedings of the 31st Conference on Artificial Intelligence, AAAI, San Francisco, California, USA., 2017, pp. 3216–3222.
[20] R. Caruana, Multitask learning, Machine Learning 28 (1) (1997) 41–75.
[21] M. P. Kumar, B. Packer, D. Koller, Self-paced learning for latent variable models, in: Advances in Neural Information Processing Systems 23, NIPS, Vancouver, British Columbia, Canada., 2010, pp. 1189–1197.
[22] J. Kivinen, A. J. Smola, R. C. Williamson, Online learning with kernels, in: Advances in Neural Information Processing Systems 14, NIPS, Vancouver, British Columbia, Canada, 2001, pp. 785–792.
[23] J. C. Schlimmer, R. H. Granger, Incremental learning from noisy data, Machine Learning 1 (3) (1986) 317–354.
[24] K. Ravi, V. Ravi, A survey on opinion mining and sentiment analysis: tasks, approaches and applications, Knowledge Based Systems 89 (2015) 14–46.
[25] 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, ACL, Baltimore, MD, USA, 2014, pp. 49–54.
[26] D. Tang, B. Qin, X. Feng, T. Liu, Effective lstms for target-dependent sentiment classification, in: Proceedings of the 26th International Conference on Computational Linguistics, COLING, Osaka, Japan, 2016, pp. 3298–3307.
[27] S. Wang, S. Mazumder, B. Liu, M. Zhou, Y. Chang, Target-sensitive memory networks for aspect sentiment classification, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL, Melbourne, Australia, 2018, pp. 957–967.
[28] X. Li, L. Bing, W. Lam, B. Shi, Transformation networks for target-oriented sentiment classification, in: Proc. 56th Ann. Meeting Assoc. Computational Linguistics, Volume 1: Long Papers, 2018, pp. 946–956.
[29] V. Mnih, N. Heess, A. Graves, K. Kavukcuoglu, Recurrent models of visual attention, in: Advances in Neural Information Processing Systems 27, NIPS, Montreal, Quebec, Canada, 2014, pp. 2204–2212.
[30] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, Y. Bengio, Attention-based models for speech recognition, in: Advances in Neural Information Processing Systems 28, NIPS, Montreal, Quebec, Canada, 2015, pp. 577–585.
[31] 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, EMNLP, Lisbon, Portugal, 2015, pp. 1412–1421.
[32] O. Firat, K. Cho, Y. Bengio, Multi-way, multilingual neural machine translation with a shared attention mechanism, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: human Language Technologies, NAACL HLT, San Diego California, USA, 2016, pp. 866–875.
[33] X. He, D. Golub, Character-level question answering with attention, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, 2016, pp. 1598–1607.
[34] J. Cheng, S. Zhao, J. Zhang, I. King, X. Zhang, H. Wang, Aspect-level sentiment classification with HEAT (hierarchical attention) network, in: Proceedings of the 2017 ACM Conference on Information and Knowledge Management, CIKM, Singapore, 2017, pp. 97–106.
[35] B. Wang, W. Lu, Learning latent opinions for aspect-level sentiment classification, in: Proceedings of the 32nd Conference on Artificial Intelligence, AAAI, New Orleans, Louisiana, USA, 2018.
[36] D. Ma, S. Li, X. Zhang, H. Wang, Interactive attention networks for aspect-level sentiment classification, in: Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI, Melbourne, Australia, 2017, pp. 4068–4074.
[37] P. Chen, Z. Sun, L. Bing, W. Yang, Recurrent attention network on memory for aspect sentiment analysis, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP, Copenhagen, Denmark, 2017, pp. 452–461.
[38] B. Huang, Y. Ou, K. M. Carley, Aspect level sentiment classification with attention-over-attention neural networks, in: Social, Cultural, and Behavioral Modeling - 11th International Conference, SBP-BRiMS, Washington, DC, USA, 2018, pp. 197–206.
[39] R. He, W. S. Lee, H. T. Ng, D. Dahlmeier, Effective attention modeling for aspect-level sentiment classification, in: Proceedings of the 27th International Conference on Computational Linguistics, COLING, Santa Fe, New Mexico, USA, 2018, pp. 1121–1131.
[40] Q. Liu, H. Zhang, Y. Zeng, Z. Huang, Z. Wu, Content attention model for aspect based sentiment analysis, in: Proceedings of the 2018 Web Conference, WWW, Lyon, France, 2018, pp. 1023–1032.
[41] F. Fan, Y. Feng, D. Zhao, Multi-grained attention network for aspect-level sentiment classification, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP, Brussels, Belgium, 2018, pp. 3433–3442.
[42] C. J. Hutto, E. Gilbert, VADER: a parsimonious rule-based model for sentiment analysis of social media text, in: Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM, Ann Arbor, Michigan, USA, 2014.
[43] X. Ding, B. Liu, P. S. Yu, A holistic lexicon-based approach to opinion mining, in: Proceedings of the International Conference on Web Search and Web Data Mining, WSDM, Palo Alto, California, USA, 2008, pp. 231–240.
[44] L. Jia, C. T. Yu, W. Meng, The effect of negation on sentiment analysis and retrieval effectiveness, in: Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM, Hong Kong, China, 2009, pp. 1827–1830.
[45] A. Hogenboom, P. van Iterson, B. Heerschop, F. Frasincar, U. Kaymak, Determining negation scope and strength in sentiment analysis, in: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Anchorage, Alaska, USA, 2011, pp. 2589–2594.
[46] D. Chen, C. D. Manning, A fast and accurate dependency parser using neural networks, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP, Doha, Qatar, 2014, pp. 740–750.

[47] S. Abbasi Moghaddam, Aspect-based opinion mining in online reviews, Ph.D. Thesis, Simon Fraser University, 2013.
[48] X. Zhou, Y. Chen, D. Z. Wang, Archimedesone: query processing over probabilistic knowledge bases, Proceedings of the VLDB Endowment 9 (13) (2016) 1461–1464.
[49] J. Yang, D. Xu, Evidential reasoning rule for evidence combination, Artificial Intelligence 205 (2013) 1–29.
[50] K. Li, X. Zhou, D. Z. Wang, C. E. Grant, A. Dobra, C. Dudley, In-database batch and query-time inference over probabilistic graphical models using UDA-GIST, The VLDB Journal 26 (2) (2017) 177–201.
[51] T. Álvarez-López, J. Juncal-Martínez, et al., GTI at semeval-2016 task 5: SVM and CRF for aspect detection and unsupervised aspect-based sentiment analysis, in: Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT, San Diego, CA, USA, 2016, pp. 306–311.
[52] M. Pontiki, D. Galanis, et al., Semeval-2015 task 12: aspect based sentiment analysis, in: Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT, Denver, Colorado, USA, 2015, pp. 486–495.
[53] M. Pontiki, D. Galanis, et al., Semeval-2016 task 5: aspect based sentiment analysis, in: Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT, San Diego, CA, USA, 2016, pp. 19–30.