Attention-Based Deep Distance Metric Learning for Aspect-Phrase Grouping

Shufeng Xiong
School of Computer, Wuhan University, Wuhan, P.R.China
Pingdingshan University, Pingdingshan, R.R.China
Donghong Ji
School of Computer, Wuhan University, Wuhan, P.R.China

Abstract

Aspect-phrase grouping is an important task for aspect finding in aspect-level sentiment analysis and it is a challenging problem due to polysemy and its context dependency. In this paper we propose an Attention-based Deep Distance Metric Learning (ADDML) method, which is more beneficial for clustering by considering aspect-phrase representation as well as context representation and their combination. Firstly, we feed word embeddings of aspect-phrases and its contexts into an attention-based network to learn feature representation of contexts. Then, both of aspect-phrase embedding and context embedding as the input of a multi-layer perceptron which is used to learn deep feature subspace, under which the distance of each intra-grouped pair is smaller and that of each inter-grouped pair is bigger, respectively. After obtaining the learned representations, we use K-means to cluster them. Our extensive experimental study on four domain review datasets shows that the proposed method outperforms baseline methods.

1 Introduction

The task of aspect phrase-grouping is to group those phrases which are referring to the same aspect(product feature) in web reviews into the same cluster. There is considerable interest in automatic aspect phrase grouping, both as an end in itself and as an intermediate step in aspect-level sentiment analysis [Hu and Liu, 2004; Pang and Lee, 2008; Liu, 2012; Fang et al., 2013; Liu et al., 2014].

Aspect-phrase Grouping is an important and necessary work for aspect-level sentiment analysis. For aspect-level sentiment analysis, it is a necessary step to find aspects from the corpus. However, people can use different words/phrases to refer to the same aspect in reviews. An aspect is the name of a feature of the product, while an aspect-phrase is a word or phrase that actually appears in a sentence to indicate the aspect. For example, “picture quality” could have some other expressions such as “photo”, “image” and “picture”. All of the aspect-phrases in a group indicate the same aspect. Therefore, aspect-phrase grouping is a necessary follow-up work after identifying aspect-phrase. In this paper, we assume that all aspect-phrases have been identified by the existing extracting methods [Jin et al., 2009; Mei et al., 2007; Kobayashi et al., 2007; Kim and Hovy, 2006; Ku et al., 2006; Hu and Liu, 2004], and we focus on grouping domain synonymous aspect-phrases.

Most of existing work employed unsupervised or semi-supervised method which exploited lexical similarity from semantic dictionary as well as context environment [Zhao et al., 2014; Zhai et al., 2011; Guo et al., 2009]. These methods achieved performance, but there are still some limitations. Firstly, their instance representation method cannot effectively express the fine-grained semantic information. In these methods, the context environment is formed by aggregating related sentences which mention the same aspect-
phrase. Thereafter, they represent aspect-phrase and context environment using bag-of-words (BoW) separately. Then both of aspect-phrase and context environment are integrated into a learning framework. Although BoW method has been shown to be effective and yield relatively high performance in many tasks of NLP. But for this task, aspect-phrase grouping, it only represents the word itself and the context them self separately which cannot preserve the accurate semantic similarities [Xu et al., 2015]. Moreover, these two parts are shallow semantic and it must do an extra combination operation for grouping task. It is attractive to consider a representation method, in which it learns a combination of aspect-phrase and its context and can preserve the semantic information. And the method is as independent from dictionary and the length of context as possible, because of there are different context lengths for different aspect-phrases.

Secondly, their distance metric method is based on the original sample representation space whose performance depends heavily on the appropriate feature representation of samples. The distance metric is an important factor for clustering algorithm which is used to divide samples into different groups. But the original sample representation is not necessarily linear separable. Yet how to design an effective model to automatically learn a distance metric is an open question.

With the recent revival of interest in Deep Neural Network (DNN), many researchers have concentrated on using Deep Learning to tackle traditional machine leaning problem. Especially, recent advances in word embedding induction methods [Pennington et al., 2014; Mikolov et al., 2013; Collobert and Weston, 2008] have benefited researches. They have enabled more effective training of DNN by providing compact input representations of the words.

Motivated by the recent success of Deep Learning, in this paper, we propose an attention-based deep distance metric learning (ADDML) method for aspect-phrase grouping. An overall architecture of the proposed method is illustrated in Figure 1. In our framework, the network structure is a parallel form. The two networks have the same parameters, each network has two main parts: (1) an attention-based semantic composition model, and (2) a nonlinear transformation model via MultiLayer Perceptron (MLP). With attention-based semantic composition model, we can naturally integrate each aspect-phrase and its variety length context into a fixed length vector. With nonlinear transformation model, we can learn a set of hierarchical nonlinear transformations to project sample pair into other feature subspace, under which the distance of each intra-grouped pair is less than a smaller threshold and that of each inter-grouped pair is larger, respectively. After obtaining the learned feature subspace, traditional K-means algorithm is employed to cluster aspect-phrases.

In summary, the main contributions of this paper are summarized as follows:

1. We explore a new attention-based method to learn the combination of aspect-phrases and their variety length context, with the help of word embedding to solve aspect-phrase grouping task.
2. We attempt to use a deep metric learning method to learn feature subspace for aspect-phrase grouping task through distance supervise, and our approach only needs a little tagged aspect-phrase pairs with the sentences in which they appeared.
3. Through experiments on four datasets, we demonstrate the effectiveness of our model.

2 Methodology

We now proceed with a discussion of our ADDML method for aspect-phrase grouping. According to the discussion in Section 1 we need to solve two main problems: (1) to express the fine-grained semantic information with a fixed length vector which can naturally combine aspect-phrase and its context and preserve the accurate semantic, and (2) to provide nonlinear transformations to learn a feature subspace under which the distance of each intra-grouped pair is smaller and that of each inter-grouped pair is larger, respectively. We firstly discuss the attention-based semantic composition model, then describe MLP-based nonlinear transformation model, the subsequent training and clustering methods.

2.1 Attention-Based Semantic Composition Model

In this model, our purpose is learning the semantic representation of the context of each aspect-phrase. Actually, the meaning of a longer expression (e.g. a context or a sentence) comes from the meanings of its words and the rules used to combine them. Two state-of-the art methods for modelling semantic representation of sentences (context) are convolutional neural network (CNN) [Kalchbrenner et al., 2014; Kim, 2014; Tang et al., 2015], and recursive neural network (RNN) [Socher et al., 2013]. Because of RNN relies on parse tree, CNN seems to be a proper choice for modelling aspect-phrase context. However, in our task, the context may be shared by more than one aspect-phrase. For example, in sentence: “the picture is clear and sharp and the sound is good.” it mentions two aspect-phrases “picture” and “sound”. Despite its computational efficiency, CNN fails to learn the difference among the shared context by different aspect-phrases. Unfortunately, in product review, one context fragment commonly mentions more than one aspect-phrase. To address this problem, we develop an attention-based semantic composition model, which considers contextual words depending on its different weight score to the aspect-phrase. In Section 3 we compare our method with CNN and the result shows the effectiveness of our method.

Actually, attention-based method has been successfully used in some neural network model [Rush et al., 2015; Ling et al., 2015; Luong et al., 2015]. We firstly propose to apply it to aspect-phrase grouping task. This model is used to transform context segment to a powerful representation. Particularly, given each word vector $e_i$, which is projected into a word embedding matrix $L_w \in \mathbb{R}^{d \times |V|}$, where $d$ is the dimension of word vector and $|V|$ is the size of word vocabulary. Let $c = \{e_i | e_i \in \mathbb{R}^{d \times 1}\}$ denote the set of input $n$
words in context, where \( l \) is the dimension of the original context segment. We employ a linear layer to combine the original context vector \( c \) and attentional weight \( a \) to produce an attentional context representation as follows:

\[
\hat{c} = f_a(c, a)
\]  

(1)

where \( f_a \) is a weighted average function. The idea of attention model is to give different weight for different words in context when deriving the context vector \( \hat{c} \). The weight \( a \in \mathbb{R}^{n \times 1} \) is a variable-length attention vector, whose size is equal to the number of words in context, is computed as follows:

\[
a(e_i) = \frac{\exp(\text{score}(e_i, p))}{\sum_i \exp(\text{score}(e_i, p))}
\]  

(2)

where \( \text{score}(e_i, p) = W_a[e_i; p] \) and \( W_a \in \mathbb{R}^{2 \times d \times 1} \) is the parameter need to be learn. Although the length of context is variety, in our model we can use the fixed-length \( W_a \) parameter to weight the importance of each word \( e_i \) for its corresponding aspect-phrase \( p \).

### 2.2 MLP-Based Nonlinear Transformation Model

After getting the attentional context \( \hat{c} \), we employ a MLP-based nonlinear transformation model to learn a feature subspace. In the new feature subspace, the distance of each inter-grouped aspect-phrase pair is larger and of each intra-grouped pair is smaller respectively. Although the attentional context \( \hat{c} \) is weighted according to its aspect-phrase, the aspect-phrase \( p \) is still a necessary information for grouping. Therefore, we concat \( \hat{c} \) and \( p \) to produce a vector \( x \) as the input of MLP.

Our model is based on a variant of Mahalanobis distance metric learning method which is proposed for face verification in the wild [Hu et al., 2014]. The problem is formulated as follows. Given a training set \( X = \{x_i | x_i \in \mathbb{R}^{d \times 1}\}_{i=1,2,...,m} \), where \( x_i \) is the \( i \)th training sample and \( m \) is the size of training set. The learning method aims to seek a linear transformation \( W \), under which the distance between any two samples \( x_i \) and \( x_j \) can be computed as:

\[
d_w(x_i, x_j) = \|Wx_i - Wx_j\|_2
\]  

(3)

where \( W \) is an alternative of the covariance matrix \( M \) in Mahalanobis distance. \( M \) can be decomposed by as follows:

\[
M = W^TW
\]  

(4)

Actually, Eq.(3) is the Euclidean distance of two samples in the linear transformed space as well as the Mahalanobis distance in the original space. The transformation \( Wx \) can be replaced with a generalized function \( g \). When \( g \) is a nonlinear function, we get the nonlinear transformation form of Mahalanobis distance. The same as [Hu et al., 2014], in our model we use the squared Euclidean distance as follows:

\[
d_g^2(x_i, x_j) = \|g(x_i) - g(x_j)\|_2^2
\]  

(5)

As shown in Fig.(1), we use hierarchical nonlinear mappings to project the sample to feature subspace. The framework is a parallel MLP networks, the top layers output the final representation of samples when feeding sample pair \( x_i \) and \( x_j \). Assume there are \( M \) layers in the designed network, and \( k^{(m)} \) units in the \( m \)th layer, where \( m = 1, 2, ..., M \).

For a given aspect-phrase sample \( x \), the output of the first layer is computed as \( h^{(1)} = f_a(W^{(1)}x + b^{(1)}) \in \mathbb{R}^{k^{(2)}} \), where the weight matrix \( W^{(1)} \in \mathbb{R}^{k^{(2)} \times k^{(1)}} \) can be seen as a nonlinear projection transformation, \( h^{(1)} \in \mathbb{R}^{k^{(2)}} \) is a bias vector, and \( f_a : \mathbb{R} \rightarrow \mathbb{R} \) is a nonlinear activation function. Subsequently, the output of the first layer \( h^{(1)} \) is used as the input of the second layer. In the same way, the output of the second layer is \( h^{(2)} = f_a(W^{(2)}h^{(1)} + b^{(2)}) \in \mathbb{R}^{k^{(3)}} \), where \( W^{(2)} \in \mathbb{R}^{k^{(3)} \times k^{(2)}} \), \( b^{(2)} \in \mathbb{R}^{k^{(3)}} \) and \( f_a \) are the projection matrix, bias and nonlinear activation function of the second layer, respectively. Finally, the output of the most top layer is calculated as follows:

\[
h^{(M)} = f_a(W^{(M)}h^{(M-1)} + b^{(M)}) \in \mathbb{R}^L
\]  

(6)

where \( L \) is the dimension of the output vector of the MLP network.

Given a pair of aspect-phrase samples \( x_i \) and \( x_j \), we let \( g(x_i) = h_i^{(M)} \) and \( g(x_j) = h_j^{(M)} \). Therefore the function \( g \) represents a hierarchical nonlinear transformation, in which the sample pair passed through the \( M \)-layer deep network and are mapped into a feature subspace. By Eq.(5), we can measure the distance between the sample pair in the new feature subspace.

The ultimate goal of the MLP-based nonlinear transformation model is to make the distance more effective for grouping the samples. To achieve this, we use a large margin framework to restrict the distance as proposed in [Mignon and Junie, 2012]. In this framework, the intra-grouped sample pair are used as positive instance and the inter-grouped sample are used as negative instance. The distance \( d_g^2(x_i, x_j) \) of positive instance \( (i,j) \) is less than a smaller threshold \( t_1 \) and that of negative instance \( (i,j) \) is larger than a larger threshold \( t_2 \), where the label \( l_{ij} \) denotes the similarity or dissimilarity between a sample pair \( x_i \) and \( x_j \), and \( t_2 > t_1 \). This constraint can be formed as follows:

\[
l_{ij}(t - d_g^2(x_i, x_j)) > 1
\]  

(7)

where \( t_1 = t - 1 \) and \( t_2 = t + 1 \), and \( t > 1 \). Eq. (7) enforce the margin between \( d_g^2(x_i, x_j) \) and \( t \) is larger than \( l \).

### 2.3 Model Training

During the training phase, each pair must satisfy the constraint in Eq. (7). Let \( \omega = 1 - l_{ij}(t - d_g^2(x_i, x_j)) \), it can be encoded into the minimization of the objective function:

\[
J = \frac{1}{2} \sum_{i,j} \sigma(\omega) + \frac{\lambda}{2} \sum_{m=1}^{M} (\|W^{(m)}\|_F^2 + \|b^{(m)}\|_2^2)
\]  

(8)

where \( \sigma(\omega) = \frac{1}{\beta} \log(1 + \exp(\beta \omega)) \) is the generalized logistic loss function [Mignon and Junie, 2012], which is a smooth approximation of the hinge loss \( E(z) = \max(0, z) \); \( \beta \) is the sharpness parameter, \( \lambda \) is a regularization parameter and \( \|W\|_F^2 \) represents the Frobenius norm of matrix \( W \).
The problem in Eq. (8) is thus solved using a stochastic sub-gradient descent scheme. The gradient of the objective function $J$ with respect to the parameters $W^{(m)}$ and $b^{(m)}$ can be calculated as follows:

$$\frac{\partial J}{\partial W^{(m)}} = \sum_{i,j} (\delta_{ij}^{(m)} h_{i}^{(m-1)T} + \delta_{ji}^{(m)} h_{j}^{(m-1)T}) + \lambda W^{(m)}$$  

$$\frac{\partial J}{\partial b^{(m)}} = \sum_{i,j} (\delta_{ij}^{(m)} + \delta_{ji}^{(m)}) + \lambda b^{(m)}$$

(9) (10)

where $h_{i}^{(0)} = x_{i}$ and $h_{j}^{(0)} = x_{j}$, and the residuals can be computed as:

$$\delta_{ij}^{(M)} = \sigma'(w)l_{ij}(h_{i}^{(M)} - h_{j}^{(M)}) \circ f_{a}'(z_{i}^{(M)})$$  

$$\delta_{ji}^{(M)} = \sigma'(w)l_{ij}(h_{j}^{(M)} - h_{i}^{(M)}) \circ f_{a}'(z_{j}^{(M)})$$

(11) (12)

$$\delta_{ij}^{(m)} = (W^{(m+1)}T \delta_{ij}^{(m+1)}) \circ f_{a}'(z_{i}^{(m)})$$  

$$\delta_{ji}^{(m)} = (W^{(m+1)}T \delta_{ji}^{(m+1)}) \circ f_{a}'(z_{j}^{(m)})$$

(13) (14)

where $z_{i}^{(m)}$ is defined as follows:

$$z_{i}^{(m)} = W^{(m)}h_{i}^{(m-1)} + b^{(m)}$$

(15)

We train the network by back-propagation and perform the following gradient descent algorithm to update parameters:

$$W^{(m)} = W^{(m)} - \mu \frac{\partial J}{\partial W^{(m)}}$$

$$b^{(m)} = b^{(m)} - \mu \frac{\partial J}{\partial b^{(m)}}$$

(16) (17)

where $\mu$ is the learning rate. In attention-based semantic composition model, $W_{a}$ is actually a linear transformation parameter which is used to learn attention weight $\alpha$, it can be calculated like the parameter $W^{(m)}$.

We set the dimension of word vector as 200, the output length of MLP as 50. The parameters of linear layer are initialized using normalized initialization [Glorot and Yoshua, 2010]. We train a three layers MLP and employ dropout with 50% rate to the hidden layer. We choose the $\text{tanh}$ as the activation function. The threshold $t$, the learning rate $\mu$ and regularization parameter $\lambda$ are empirical set as 3, 0.03 and 0.002 for all experiments, respectively.

### 2.4 K-means for Clustering

With the given texts, we firstly employ the parallel deep neural networks to learn the semantic representation $h^{(M)}$, and then utilize traditional K-means algorithm to perform clustering.

### 3 Experiments

We validate the proposed approach on four different domain datasets. In the following, we firstly present the datasets and the evaluation measures and then present our results.

#### 3.1 Data Preparation

Four product domains of customer reviews from Customer Review Datasets (CRD) [Hu and Liu, 2004]: digital camera (DC), DVD player, MP3 player (MP3) and cell phone (PHONE) are employed to evaluate our proposed approach. The aspect label of each aspect-phrase is hand-annotated.

For obtaining the training sample pair set of the parallel networks, we firstly randomly select 3 groups as the seed training set. We utilize each sentence with its mentioned aspect-phrase as the original training data. We then combine aspect-phrase and its related sentences to form sample set which is the actual input to the parallel networks. For example, given a aspect-phrase $p_{1}$ and a sentence set $S_{1} = \{s_{1,1}, s_{1,2}, ..., s_{1,m}\}$ in which there are $m$ sentences and $s_{1,i}$ mentions $p_{1}$, we can construct $m$ samples as $\{p_{1} \cup s_{1,1}, p_{1} \cup s_{1,2}, ..., p_{1} \cup s_{1,m}\}$. The group label of the sample is the same as its original aspect-phrase, e.g. when $p_{1}$ belong to group 1, then all of $p_{1} \cup s_{1,i}$ have the group label 1. Thereafter, we find all non-repetitive item pair in the same group as the positive sample pair set, assume there are $M$ samples in group 1, the number of sample pair is $\binom{M}{2}$. For balancing the training set, we find the same number of negative sample pair in which two samples are selected from different groups. For testing data, we still aggregate all the related sentences to form the context environment. The statistics are described in Table I.

![Table 1: Statistics of the review corpus. # donotes the size](image)

|          | DC  | DVD | MP3 | PHONE |
|----------|-----|-----|-----|-------|
| #Sentences| 330 | 247 | 581 | 231   |
| #Aspect-phrases| 141 | 109 | 183 | 102   |
| #Aspects   | 14  | 10  | 10  | 12    |
| #Pairs     | 19163 | 11211 | 64945 | 8855  |

#### 3.2 Pre-trained Word Vectors

We use the publicly available Glove tools to train word embeddings, and the most parameters are set as same as [Pennington et al., 2014] to train word vectors. Because of the review corpus is too small for learning word embeddings, we use Amazon Product Review Data [Jindal and Liu, 2008] as the auxiliary training corpus. The vectors have dimension of 200 and the words not present in the set of pre-trained words are initialized randomly.

#### 3.3 Evaluation Measures

Since the problem of grouping aspect-phrase is a clustering task, four common measures for clustering algorithm are used to performance evaluation: Purity, Entropy, Normalized Mutual Information (NMI) and Rand Index (RI).

#### 3.4 Baseline Methods and Settings

The proposed ADDML method is compared with a number of existing methods, which can be classified as state-of-the-art methods, word embedding composite methods and word embedding-based distance learning methods.
In the state-of-the-art series, all of them exploit the labelled data which is generated using sharing-word constraint and lexical similarity based on WordNet except Kmeans and CC-Kmeans:

**Kmeans:** It is the most popular clustering algorithm based on distributional similarity with cosine as the similarity measure and BoW as the feature representation.

**DF-LDA:** It is a combination of Dirichlet Forest Prior and LDA model, in which it can encode domain knowledge (the label) into the prior on topic-word multinomials [Andrzejewski et al., 2009]. The code is available in author’s website.

**L-EM:** This is a state-of-the-art semi-supervised method for clustering aspect-phrases [Zhai et al., 2011a]. L-EM employed lexical knowledge to provide a better initialization for EM. We implement this method by ourself code.

**CC-Kmeans:** It is proposed in [Xiong and Ji, 2015], in which it encodes the capacity limitation as constraint and proposes a capacity constrained K-means algorithm to cluster aspect-phrases. We use the code from the author.

In the word embedding composite methods series, these methods employ different composite strategies to form the sample vector, respectively. The clustering method is Kmeans with cosine distance in which word embedding is used as feature vector.

**AVG/MIN/MAX+MLP:** They use the average/minimize/maximize value of all the context word vectors in each dimension as the context vector \( \tilde{c} \), respectively. And then, aspect-phrase \( p \) and \( \tilde{c} \) are concatenated to form the sample vector.

**AP:** This method only uses Aspect-Phrase (AP) vector to cluster aspect-phrases.

In the word embedding-based distance learning methods, all of the methods are used to learn the vector representation, and the clustering method is Kmeans with Euclidean distance because the learning method is considering optimize the Euclidean distance. They are:

**CNN+MLP:** It uses CNN to learn the context vector \( \tilde{c} \), the other component is the same as ADDML.

**ADML:** In this method, it only uses attention-based semantic composition model and distance metric learning. In other words, it does not use non-linear transformation to learn feature subspace compared with ADDML.

Since all methods based on Kmeans depend on the random initiation, we use the average results of 10 runs as the final result. For L-EM, we use the same parameter settings with the original paper.

### 3.5 Evaluation Results

Now, we present and compare the results of ADDML and the 10 baseline methods based on 4 domains. All the results are shown in Table 2, Table 3. \( \text{avg} \) represents the average result of the 4 domains. The results are separated into three groups according to categories of the baseline methods. For **Entropy**, the smaller value is the better, but for **Purity, NMI and RI**, the larger the better. We can see that our approach outperforms baseline methods on the average result of all domains. In addition, we make the following observations:

- From the first group, we can see that L-EM and CC-Kmeans perform better than other methods and all of them are weakly than ADDML. In this group, the methods exploit external knowledge and constraint can achieve better performance but still not efficient.
- From the second group, all of the methods employ the word embedding to represent and composite the word semantic. Yet these methods achieve uneven result attribute to different composite strategies. The AVG method performs better than others in overall result, in which it average the semantic of each words in the context. The average operate is a common used approach in many word embedding-based methods, such as CNN, and achieves better performance. However, it still falls behind our ADDML according to its task-independent property. Especially, AP method obtains best result in MP3 domain. We argue that the main reason is the most aspect-phrases themselves have fixed meanings in this domain, and so it achieves best performance by only using the aspect-phrase word vector.
- From the third group, three methods aim to learn a feature subspace via different neural network framework. CNN+MLP and ADML achieve better results in different domains respectively. For example, CNN+MLP is good in MP3 and PHONE domains but ADML is good in DC and DVD domains. We have mentioned that the aspect-phrase in MP3 domain may have fixed meanings, in other words aspect-phrase and its context have less relevance under the grouping task. Therefore, attention-based methods (ADML and ADDML) do not achieve best performance in MP3 dataset. Nevertheless, ADDML still gets better results in Purity and Entropy in MP3 domain. By using the MLP-based nonlinear transformation, ADDML performs better than ADML which only uses attention-based semantic composite vector for distance metric learning in all domains.

### 4 Related Work

Our work is related to three important research topics: aspect-level sentiment analysis, metric learning, and deep learning.

For aspect-level sentiment analysis, there are many studies on clustering aspect-phrases. There are some topic-model-based approaches that works joint extract aspect-phrases and group them at the same time [Chen et al., 2013; Moghaddam and Ester, 2012; Lu et al., 2011; Jo and Oh, 2011; Zhao et al., 2010; Lin and He, 2009; Titov and Ryan, 2008]. Those methods tend to discover coarse-grained and grouped aspect-phrases but not specific opinionated aspect-phrase themselves. In addition, [Zhai et al., 2011a] showed that it has not perform well even considering pre-existing knowledge. And some other work focus on grouping aspect-phrases. [Guo et al., 2009] grouped aspect-phrases using multi-level
LaSA method which exploited the virtual context documents and semantic structure of aspect-phrase. [Zhai et al., 2011b; Zhai et al., 2010] used an EM-based semi-supervised learning method for clustering aspect-phrases in which the lexical knowledge is used to provide a better initialization for EM. [Zhao et al., 2014] proposed a framework of Posterior Regularization to cluster aspect-phrases in which it formalizes sentiment distribution consistency as soft constraint. This method requests a special semi-structured customer reviews to estimate the sentiment distribution. Therefore, our method does not compared with it.

Our work is also related to metric learning algorithms. Most of them have been successfully applied to address the problem of face verification [Shi et al., 2015; Ding et al., 2015; Yi et al., 2014; Cai et al., 2012; Hu et al., 2014]. The common objective of these methods is to learn a better distance metric so that the distance between positive pair is smaller and the distance between negative pair is larger. However, these methods did not perform the nonlinear transformation. In addition, [Hu et al., 2014] employed a MLP-based nonlinear transformation, but its input was the given image descriptor which can be directly concatenated to form feature vectors. In this paper, we introduce this method to aspect-phrase grouping task, and provide an extra attention-based semantic composite model to obtain feature vectors based on word vectors of aspect-phrase and its context.

In our work, the basis is word embedding. The word embedding is also called distributed word representation, it is typically induced using neural language models, which use neural networks as the underlying predictive model. [Collobert and Weston, 2008; Mnih and Hinton, 2007; Mikolov et al., 2013] presented different models to improve the performance and efficiency. Especially, [Pennington et al., 2014] proposed Glove, a new global log-bilinear regression model for the unsupervised learning of word representations. Based on word embedding, neural networks can capture true meaningful syntactic and semantic regularities, most of deep learning neural networks can capture true meaningful syntactic regularities, most of deep learning methods [Kalchbrenner et al., 2014; Rush et al., 2015; Kim, 2014; Luong et al., 2015; Socher et al., 2013; Xu et al., 2015] applied in many NLP tasks. In this paper, we explore the idea of attention in [Rush et al., 2015; Luong et al., 2015; Ling et al., 2015] and incorporate with MLP network to tackle aspect-phrase grouping problem.

### 5 Conclusion
This paper studies the problem of product aspect-phrase grouping for aspect-level sentiment analysis. For this grouping task, this paper explores a novel deep neural network...
framework. We firstly employ attention-based semantic composite model to obtain the weighted context vector. Then, we concatenate aspect-phrase vector and its context vector and feed them into MLP-based nonlinear transformation model to learn a feature subspace. Finally, we use a large margin constraint in the objective function, which can ensure the distance of each inter-grouped sample pair is larger and of each intra-grouped pair is smaller in the new feature subspace respectively. By using the learnt feature representation as input in K-means algorithm, experiments show that our approach is superior to other baselines. How to apply our ADDML method to other NLP application such as short text clustering and sentence similarity measure is an interesting direction of future work.

References

[Andrzejewski et al., 2009] D. Andrzejewski, X. Zhu, and M. Craven. Incorporating domain knowledge into topic modeling via dirichlet forest priors. *Proc Int Conf Mach Learn*, 382(26):25–32, 2009.

[Cai et al., 2012] Xinyuan Cai, Chunheng Wang, Baihua Xiao, Xue Chen, and Ji Zhou. Deep nonlinear metric learning with independent subspace analysis for face verification. In *Proceedings of the 20th ACM International Conference on Multimedia, MM ’12*, pages 749–752, New York, NY, USA, 2012.

[Chen et al., 2013] Zhiyuan Chen, Arjun Mukherjee, Bing Liu, Meichun Hsu, Mal Castellanos, and Riddhiman Ghosh. Exploiting domain knowledge in aspect extraction. In *EMNLP*, EMNLP, pages 1655–1667. ACL, 2013.

[Collobert and Weston, 2008] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *ICML*, ICML, pages 160–167, New York, NY, USA, 2008.

[Ding et al., 2015] Shengyong Ding, Liang Lin, Guangrun Wang, and Hongyang Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 2015.

[Fang et al., 2013] Lei Fang, Minlie Huang, and Xiaoyan Zhu. Exploring weakly supervised latent sentiment explanations for aspect-level review analysis. In *Proceedings of the 22nd ACM international conference on Conference on information and knowledge management*, CIKM, pages 1057–1066. ACM, 2013.

[Glorot and Yoshua, 2010] Xavier Glorot and Bengio Yoshua. Understanding the difficulty of training deep feedforward neural networks. In *International conference on artificial intelligence and statistics*, International conference on artificial intelligence and statistics, pages 249–256, 2010.

[Guo et al., 2009] Honglei Guo, Huijia Zhu, Zhili Guo, Xiaoxun Zhang, and Zhong Su. Product feature categorization with multilevel latent semantic association. In *CIKM*, CIKM, pages 1087–1096. ACM, 2009.

[Hu and Liu, 2004] Mingqing Hu and Bing Liu. Mining and summarizing customer reviews. *KDD*, pages 168–177. ACM, 2004.

[Hu et al., 2014] Junlin Hu, Jiwen Lu, and Yap-Peng Tan. Discriminative deep metric learning for face verification in the wild. In *cvpr*, cvpr, pages 1875 – 1882. IEEE, 2014.

[Jin et al., 2009] Wei Jin, Hung Hay Ho, and Rohini K. Srinhari. Opinionminer: A novel machine learning system for web opinion mining and extraction. In *KDD ’09*, KDD ’09, pages 1195–1204, New York, NY, USA, 2009. ACM.

[Jindal and Liu, 2008] Nitin Jindal and Bing Liu. Opinion spam and analysis. In *WSDM*, WSDM, pages 219–230. ACM, 2008.

[Jo and Oh, 2011] Yohan Jo and Alice H. Oh. Aspect and sentiment unification model for online review analysis. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM ’11, pages 815–824, New York, NY, USA, 2011. ACM.

[Kalchbrenner et al., 2014] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A convolutional neural network for modelling sentences. In *ACL*, ACL, 2014.

[Kim and Hovy, 2006] Soo-Min Kim and Eduard Hovy. Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proc. ACL Workshop on Sentiment and Subjectivity in Text*, Proc. ACL Workshop on Sentiment and Subjectivity in Text, pages 1–8, Sydney, Australia, 2006. Association for Computational Linguistics.

[Kim, 2014] Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, EMNLP, pages 1746–1751, Doha, Qatar, 2014.

[Kobayashi et al., 2007] Nozomi Kobayashi, Kentaro Inui, and Yuji Matsumoto. Extracting aspect-evaluation and aspect-of relations in opinion mining. *Proc. AAAI-CAAW*, 2007.

[Ku et al., 2006] Lun-Wei Ku, Yu-Ting Liang, and Hsin-Hsi Chen. Opinion extraction, summarization and tracking in news and blog corpora. In *Proc. AAAI-CAAW*, volume 100107 of *Proc. AAAI-CAAW*, 2006.

[Lin and He, 2009] Chenghua Lin and Yulan He. Joint sentiment/topic model for sentiment analysis. In *Proceedings of CIKM*, CIKM, pages 375–384, Hong Kong, China, 2009. ACM. 1646003.

[Ling et al., 2015] Wang Ling, Yulia Tsvetkov, Silvio Amir, Ramon Fernandez, Chris Dyer, Alan W. Black, Isabel Trancoso, and Chu-Cheng Lin. Not all contexts are created equal: Better word representations with variable attention. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, EMNLP, pages 1367–1372, Lisbon, Portugal, 2015.
In *Proc. EMNLP*, Proc. EMNLP, pages 1614–1623. Association for Computational Linguistics, 2014.