

# Distance Metric Learning for Aspect Phrase Grouping

Shufeng Xiong<sup>1,3</sup>, Yue Zhang<sup>2</sup>, Donghong Ji<sup>1\*</sup>, Yinxia Lou<sup>1</sup>

<sup>1</sup>Computer School, Wuhan University, Wuhan, China

<sup>2</sup>Singapore University of Technology and Design, Singapore

<sup>3</sup>Pingdingshan University, Pingdingshan, China

{xsf, dhji}@whu.edu.cn

yue\_zhang@sutd.edu.sg

## Abstract

Aspect phrase grouping is an important task in aspect-level sentiment analysis. It is a challenging problem due to polysemy and context dependency. We propose an Attention-based Deep Distance Metric Learning (ADDML) method, by considering aspect phrase representation as well as context representation. First, leveraging the characteristics of the review text, we automatically generate aspect phrase sample pairs for distant supervision. Second, we feed word embeddings of aspect phrases and their contexts into an attention-based neural network to learn feature representation of contexts. Both aspect phrase embedding and context embedding are used to learn a deep feature subspace for measure the distances between aspect phrases for K-means clustering. Experiments on four review datasets show that the proposed method outperforms state-of-the-art strong baseline methods.

## 1 Introduction

For aspect-level sentiment analysis (Hu and Liu, 2004; Pang and Lee, 2008), aspect identification from the corpus is a necessary step. Here *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. Different aspect phrases can be used to describe the same aspect. For example, “picture quality” could be referred to “photo”, “image” and “picture”. All aspect phrases in the same group indicate the same aspect. In this paper, we assume that all aspect phrases have been identified by using existing methods (Jin et al., 2009; Kobayashi et al., 2007; Kim and Hovy, 2006), and focus on grouping domain synonymous aspect phrases.

Most existing work employed unsupervised methods, exploiting lexical similarity from semantic dictionary as well as context environments (Zhao et al., 2014; Zhai et al., 2011a; Guo et al., 2009). The context for an aspect phrase is formed by aggregating related sentences that mention the same aspect phrase. Thereafter, aspect phrase and context environment are represented using bag-of-word (BoW) models separately, and integrated into a unified learning framework.

One limitation of the existing methods is that they do not model the interaction between aspect phrases and their contexts explicitly. For example, in the review “the picture is clear, bright and sharp and the sound is good”, the words “clear”, “bright” and “sharp” are related to the aspect phrase “picture”, while the word “good” is related to the aspect phrase “sound”. By the traditional model, these words are not differentiated when they are taken for the context, thereby causing noise in the grouping of the two aspect phrases.

To address this issue, we propose a novel neural network structure that automatically learns the relative importance of each context word with respect to a target/aspect phrase, by leveraging an attention model (Luong et al., 2015; Rush et al., 2015; Ling et al., 2015). As shown in Figure 1, given a sentence that contains an aspect phrase, we use a neural network to find a vector representation of the aspect phrase and its context. For the grouping of a certain aspect phrase, we concatenate all the occurrences of the aspect phrase in a corpus to find its vector form. Thus, the problem of aspect phrase grouping is

\*corresponding author

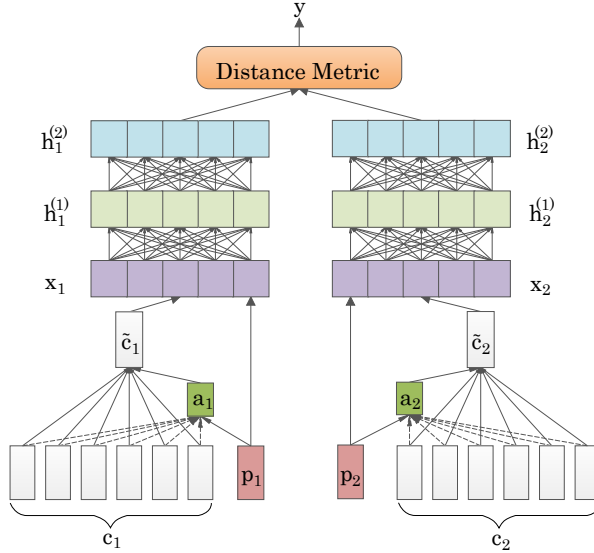


Figure 1: Architecture of the proposed method. For a given pair of aspect phrases  $p_1$  and  $p_2$ , with their contexts  $c_1$  and  $c_2$  respectively, two vectors  $x_1$  and  $x_2$  are obtained via attention-based semantic combination, and then mapped into the same feature subspace as  $h_1^{(2)}$  and  $h_2^{(2)}$ .

transformed into a clustering problem in the resulting vector space. Different from traditional methods, which leverage a bag-of-words feature space, our vector space considers not only words, but also semantic similarities between aspect phrases and contexts (Xu et al., 2015).

One challenge to the success of our method is the finding of a proper training algorithm for the neural network model. Inspired by word embedding training methods (Collobert et al., 2011; Mikolov et al., 2013), we take a negative sampling approach. In particular, we take pairs of sentences that contain the same aspect phrase as positive training examples, and pairs of sentences that contain incompatible aspect phrase as negative training examples, maximizing a score margin between positive and negative examples. Here two aspect phrases are incompatible if the distance based on a semantic lexicon is large (Faruqui et al., 2015; Yu and Dredze, 2014).

To find a better vector space representation, we add two nonlinear transformation layers, as shown in  $h^{(1)}$  and  $h^{(2)}$  in Figure 1. This method is similar to the Mahalanobis distance metric learning for face verification (Hu et al., 2014). Model training is performed by back-propagation over all neural nodes. With such vector space being learned, direct K-means clustering can be used to group aspect phrases.

Results on a standard benchmark show that our neural network significantly outperforms traditional models. The average results on 4 domains reached 0.51 (Purity) and 1.74 (Entropy), better than the previous best result (0.43 in Purity and 2.02 in Entropy).

## 2 Method

Our proposed model addresses two main problems: (1) to express fine-grained semantic information with a fixed length vector, which can naturally combine aspect phrases and their contexts, and (2) to provide nonlinear transformations to learn a feature subspace, under which the distance between each intra-group aspect phrase pair is smaller than that between each inter-group pair. We discuss an attention-based semantic composition model in Section 2.1, and then describe a Multi-Layer Perceptron (MLP)-based nonlinear transformation model in Section 2.2, which are designed for (1) and (2), respectively.

### 2.1 Attention-Based Semantic Composition

The goal of the model is to learn the semantic representation of the context of each aspect phrase. For our task, the same context is frequently shared by more than one aspect phrases. For example, in the sentence: “the **picture** is clear and sharp and the **sound** is good.”, two different aspect phrases “picture”

the	picture	is	clear	,	bright	and	sharp	and	the	sound	is	good
the	picture	is	clear	,	bright	and	sharp	and	the	sound	is	good
but	overall	this	is	a	good	camera	with	a	really	good	picture	clarity
but	overall	this	is	a	good	camera	with	a	really	good	picture	clarity

Figure 2: Visualization of weighted context by our attention model. There is a context which contains two aspect phrases (white background), where the weight of each word is taken from the developing dataset. For different aspect phrases, the words in the same context have different weights. Different background represent different weights, bigger weights corresponding to deeper background.

and “sound” are mentioned. We use an attention-based neural semantic composition model to consider contextual words based on their weight scores with respect to each aspect phrase. In particular, 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. All of  $e_i$  can be randomly initialized from a uniform distribution, and then updated during the back propagation training procedure. Alternatively, another way is using pre-trained vectors as initialization, which is learnt from text corpus with embedding learning algorithms. In our experiment, we adopt the latter strategy. Let  $c = \{e_i | e_i \in \mathbb{R}^{d \times 1}\}_{i=1,2,\dots,n}$  denote the set of  $n$  input 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:

$$\tilde{c} = f_w(c, a), \quad (1)$$

where  $f_w$  is a weighted average function. The idea is to give different weights for different words in the context when deriving the context vector  $\tilde{c}$ . 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 the context. Its value is computed as follows:

$$a(e_i) = \frac{\exp(\text{score}(e_i, p))}{\sum_{i'} \exp(\text{score}(e_{i'}, p))}, \quad (2)$$

where  $\text{score}(e_i, p) = W_a^T [e_i; p]$  and  $W_a \in \mathbb{R}^{(2*d) \times 1}$  is a model parameter to learn. Although the length of context is variable, our model uses a fixed-length  $W_a$  parameter to weight the importance of each word  $e_i$  for its corresponding aspect phrase  $p$ . This results in a fixed length vector form for each aspect phrase in a variable-size context.

## 2.2 MLP-Based Nonlinear Transformation

After obtaining the attention-based context  $\tilde{c}$ , we employ a MLP-based nonlinear transformation to learn a feature subspace for final aspect phrase grouping. Although  $\tilde{c}$  is a weighted context according to the aspect phrase, the aspect phrase  $p$  itself is still a necessary source of information for grouping. Therefore, we concatenate  $\tilde{c}$  and  $p$  to produce a vector  $x$  as the input to MLP.

Our model is based on a variant of the Mahalanobis distance metric learning method (Hu et al., 2014). The problem is formulated as follows. Given a training set  $X = \{x_i | x_i \in \mathbb{R}^{(2*d) \times 1}\}_{i=1,2,\dots,m}$ , where  $x_i$  is the  $i$ th training sample and  $m$  is the size of training set. The 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:

$$d_w(x_i, x_j) = \|Wx_i - Wx_j\|_2 \quad (3)$$

where  $W$  is an alternative of the covariance matrix  $M$  in Mahalanobis distance.  $\|A\|_2$  represents the  $L_2$  norm of the matrix  $A$ .  $M$  can be decomposed by:

$$M = W^T W, \quad (4)$$

Further, Equation (3) can be rewritten as

$$\begin{aligned}
d_w(x_i, x_j) &= \|Wx_i - Wx_j\|_2 \\
&= \sqrt{(x_i - x_j)^T W^T W (x_i - x_j)} \\
&= \sqrt{(x_i - x_j)^T M (x_i - x_j)}
\end{aligned} \tag{5}$$

Equation (5) is the typical form of Mahalanobis distance between  $x_i$  and  $x_j$ . Therefore, Equation (3) is both the Euclidean distance of two samples in the linear transformed space, and 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 obtain the nonlinear transformation form of Mahalanobis distance. Following Hu et al. (2014), we use the squared Euclidean distance in our model:

$$d_g^2(x_i, x_j) = \|g(x_i) - g(x_j)\|_2^2 \tag{6}$$

As shown in Figure 1, we use hierarchical nonlinear mappings to project the samples to a feature subspace. Assume that there are  $M$  layers in the designed network, and  $k^{(m)}$  units in the  $m$ th layer, where  $m = 1, 2, \dots, M$ . For a given aspect phrase sample  $x$ , the output of the first layer is computed as

$$h^{(1)} = f_a(W^{(1)}x + b^{(1)}) \in \mathbb{R}^{k^{(2)}}, \tag{7}$$

where the weight matrix  $W^{(1)} \in \mathbb{R}^{k^{(2)} \times k^{(1)}}$  can be seen as a linear projection transformation,  $b^{(1)} \in \mathbb{R}^{k^{(2)}}$  is a bias vector, and  $f_a : \mathbb{R} \mapsto \mathbb{R}$  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

$$h^{(2)} = f_a(W^{(2)}h^{(1)} + b^{(2)}) \in \mathbb{R}^{k^{(3)}}, \tag{8}$$

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, a bias and a nonlinear activation function of the second layer, respectively.

Finally, the output of the topmost layer is calculated as follows:

$$h^{(M)} = f_a(W^{(M)}h^{(M-1)} + b^{(M)}) \in \mathbb{R}^L, \tag{9}$$

where  $L$  is the dimension of the output vector.

Given a pair of aspect phrase samples  $x_i$  and  $x_j$ , let  $g(x_i) = h_i^{(M)}$  and  $g(x_j) = h_j^{(M)}$ . The function  $g$  represents a hierarchical nonlinear transformation, in which the aspect phrase sample pair is passed through the  $M$ -layer deep network and mapped into a feature subspace. By using Equation (6), we can measure the distance between the sample pair in the new feature subspace.

### 2.3 K-means Clustering

With a given corpus, we first employ the parallel deep neural network to learn the semantic representation  $h^{(M)}$ , and then utilize the K-means algorithm to perform clustering of  $h^{(M)}$ . During training, we sample aspect phrase pairs using the sentence as context. During testing, we concatenate all the sentences that mention the aspect phrase for clustering the aspect phrase.

### 2.4 Model Training

The ultimate goal of our model is to make the distance metric effective for grouping aspect phrase samples. To achieve this, we use a large-margin framework to restrict the distance, as proposed by Mignon and Jurie (2012). In particular, sample pairs containing the same aspect phrase are used as positive instances and sample pairs with incompatible aspect phrase are used as negative instances.

We exploit lexical similarity to obtain incompatible aspect phrases, which have low similarity in semantic lexicon. In particular, we choose WordNet as the semantic lexicon. Two aspect phrase are incompatible when the WordNet similarity between them is smaller than a threshold  $\eta$ . And the WordNet similarity is calculated by Equation (12)

$$Res(w_1, w_2) = IC(LCS(w_1, w_2)) \quad (10)$$

$$IC(w) = -\log P(w) \quad (11)$$

$$Jcn(w_1, w_2) = \frac{1}{IC(w_1) + IC(w_2) - 2 \times Res(w_1, w_2)} \quad (12)$$

where LCS (least common subsumer) is the most specific concept that is a shared ancestor of the two concepts represented by the words (Pedersen, 2010).  $P(w)$  is the probability of the concept word  $w$ . In our experiments, the threshold is set to 0.85.

Traditional methods (Zhai et al., 2010; Zhai et al., 2011b) exploit lexical knowledge to provide soft constraint for clustering aspect phrases. They assume that the aspect phrases that have high similarity in semantic lexicon, are likely to belong to the same group. In this cause, our method uses a similar assumption.

For obtaining the training data, we apply an extra sample pair generation process. The generated sample aspect phrase pairs are fed into left and right sub neural network of Figure 1, respectively. Specifically, each training sentence is utilized with its labelled aspect phrase as a gold sample. Then, we combine each aspect phrase and its related sentences to form training sample set. For example, given an aspect phrase  $p_1$  and a sentence set  $S^1 = \{s_1^1, s_2^1, \dots, s_m^1\}$ , in which there are  $m$  sentences that mention  $p_1$ , we can construct  $m$  samples  $\{p_1 \cup s_1^1, p_1 \cup s_2^1, \dots, p_1 \cup s_m^1\}$ . The group label of the sample is the same as its original aspect phrase, e.g. when  $p_1$  belongs to group 1, then all of  $p_1 \cup s_i^1$  have the group label 1.

Assuming that there are  $n$  selected aspect phrases, the number of positive sample pairs candidates is  $\binom{n}{2}$ . A negative sample pair is formed by randomly selecting incompatible aspect phrases and their contexts. For balancing the training set, we obtain the same number of negative sample pairs.

In the training objective, the distance  $d_g^2(x_i, x_j)$  of positive instances ( $l_{ij} = 1$ ) is less than a small threshold  $t_1$  and that of negative instance ( $l_{ij} = -1$ ) is higher than a large threshold  $t_2$ , where the label  $l_{ij}$  denotes the similarity or dissimilarity between a sample pair  $x_i$  and  $x_j$ , and  $t_2 > t_1$ . Let  $t > 1$ ,  $t_1 = t - 1$  and  $t_2 = t + 1$ . This constraint can be formulated as follows:

$$l_{ij}(t - d_g^2(x_i, x_j)) > 1, \quad (13)$$

Equation (13) enforces the margin between  $d_g^2(x_i, x_j)$ .

During the training phase, each aspect phrase pair must satisfy the constraint in Equation (13). Let  $\omega = 1 - l_{ij}(t - d_g^2(x_i, x_j))$ , we minimize the objective function:

$$J = \frac{1}{2} \sum_{i,j} \sigma(\omega) + \frac{\lambda}{2} \sum_{m=1}^M (\|W^{(m)}\|_F^2 + \|b^{(m)}\|_2^2) \quad (14)$$

where  $\sigma(\omega) = \frac{1}{\beta} \log(1 + \exp(\beta\omega))$  is the generalized logistic loss function (Mignon and Jurie, 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 minimization problem in Equation (14) is solved using a stochastic sub-gradient descent scheme. We train the network using back-propagation. We set the dimension of word vectors as 200, the output length of MLP as 50. The parameters of the linear layer are initialized using normalized initialization (Glorot and Yoshua, 2010). We train a three-layer MLP and employ dropout with 50% rate to the hidden layer. We choose *tanh* as the activation function. The threshold  $t$ , the regularization parameter  $\lambda$  and learning rate  $\mu$  are empirical set as 3, 0.002 and 0.03 for all experiments, respectively.

### 3 Experiments

#### 3.1 Data Preparation

We employ the datasets of Xiong and Ji (2015) to evaluate our proposed approach. The datasets are based on *Customer Review Datasets* (CRD) (Hu and Liu, 2004), including four different domains: digital camera (DC), DVD player, MP3 player (MP3) and cell phone (PHONE). We take 3 different random splits of datasets (30% train, 50% test, and 20% development). The statistics are described in Table 1.

	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

Table 1: Statistics of the review corpus. # donotes the size

#### 3.2 Pre-trained Word Vectors

We use *Glove*<sup>1</sup> tools to train word embeddings, and the training parameters are set by following Pennington et al. (2014). Because of the review corpus is too small for learning word embeddings, we use *Amazon Product Review Data* (Jindal and Liu, 2008) as one auxiliary training corpus.

#### 3.3 Evaluation Metrics

Since the problem of aspect phrase grouping is a clustering task, two common measures for clustering are used to its evaluate performance (Zhai et al., 2010) : *Purity* and *Entropy*.

**Purity** is the percentage of correct clusters that contain only data from the gold-standard partition. A large Purity reflects a better model.

**Entropy** looks at how various groups of data are distributed within each cluster. A smaller Entropy reflects a better model.

#### 3.4 Baseline Methods and Settings

The proposed ADDML method is compared with a number of methods, which include (1) the state-of-the-art methods, (2) baseline neural methods.

In the first category<sup>2</sup>, all methods except Kmeans and CC-Kmeans exploit labelled data, which is generated using sharing word constraint and lexical similarities based on WordNet<sup>3</sup>:

**Kmeans**. The most popular clustering algorithm based on distributional similarity with *cosine* as similarity measure and BoW as feature representation.

**DF-LDA**. 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)<sup>4</sup>. The code is available in author’s website<sup>5</sup>.

**L-EM**. A state-of-the-art semi-supervised method for clustering aspect phrases (Zhai et al., 2011a). It employed lexical knowledge to provide a initialization for EM. We implemented this method ourselves.

**CC-Kmeans**. It is proposed by Xiong and Ji (2015), it encodes the capacity limitation as constraint and proposes a capacity constrained K-means to cluster aspect phrases. We use the code from the author<sup>6</sup>.

<sup>1</sup><http://nlp.stanford.edu/projects/glove/>

<sup>2</sup>Because we use sample pairs but not cluster label for training, it is not possible to train a supervised classifier for testing.

<sup>3</sup>The generation of labelled data follows Zhai et al. (2011a).

<sup>4</sup>There are other LDA based methods for this task, such Constrained LDA (Zhai et al., 2010). Although Constrained LDA used two domains data form CRD dataset as core corpus, they actually crawled many other camera and phone reviews. So we are unable to compare with them by directly using their published results. Since DF-LDA is more effectiveness and suitable for our smaller datasets, we use DF-LDA as the representative of the LDA-based methods.

<sup>5</sup>[http://pages.cs.wisc.edu/~andrzej/research/df\\_lda.html](http://pages.cs.wisc.edu/~andrzej/research/df_lda.html)

<sup>6</sup><https://github.com/pdsujnow/cc-kmeans>

The word embedding composite 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** use the average/minimum/maximum value of all the context word vectors in each dimension as the context vector  $\tilde{c}$ , respectively, and then, concatenates aspect phrase  $p$  and  $\tilde{c}$  to form the sample vector.

**AP** only uses aspect phrase (AP) vector to cluster aspect phrases.

Since all the 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 Results

	Purity					Entropy				
	DC	DVD	MP3	PHONE	avg	DC	DVD	MP3	PHONE	avg
Kmeans	0.4079	0.3922	0.3509	0.3333	0.3711	2.2627	2.0056	2.2862	2.5894	2.2860
DF-LDA	0.4365	0.4362	0.3467	0.4329	0.4132	2.1355	1.9705	2.2054	2.3875	2.1747
L-EM	0.4605	0.4706	0.3333	0.4561	0.4301	2.0451	1.9145	2.2427	1.8952	2.0244
CC-Kmeans	0.4554	0.4483	0.3333	0.4353	0.4181	1.9604	1.9841	2.2897	1.8794	2.0284
AVG	0.5089	0.4483	0.3667	0.4941	0.4545	1.7203	2.1759	2.2030	1.7039	1.9508
MIN	0.4554	0.3218	0.3600	0.4118	0.3872	2.1055	2.5479	2.6598	2.2158	2.3822
MAX	0.4554	0.3563	0.3667	0.4353	0.4034	2.1230	2.4440	2.6036	2.1744	2.3363
AP	0.4196	0.4253	0.3600	0.4588	0.4159	2.1816	2.2074	2.3087	1.9946	2.1731
<b>ADDML</b>	<b>0.5658</b>	<b>0.5098</b>	<b>0.3684</b>	<b>0.6143</b>	<b>0.5146</b>	<b>1.7119</b>	<b>1.8043</b>	<b>2.1274</b>	<b>1.3282</b>	<b>1.7429</b>

Table 2: Comparison of Purity and Entropy with baselines. Our model is ADDML.

We present and compare the results of ADDML and the 8 baseline methods based on 4 domains. The results are shown in Table 2, where **avg** represents the average result of the 4 domains. The results are separated into two groups according to categories of the baseline methods. 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 the other methods. The methods that exploit external knowledge and constraint can achieve better performance. However, the proposed ADDML method outperforms all baselines by using weighted contexts as well as distance metric learning.
- From the second group, all methods employ word embeddings to represent word semantic and text composition semantic. Yet these methods achieve uneven results due to different semantic composition strategies. The neural bag-of-word AVG method performs better than the others in the overall result, in which it averages the semantics of each word in the context. The average operation is a commonly used approach in many neural methods, such as CNN (Convolution Neural Network), and achieves better performance. However, it still falls behind our ADDML method according to its task-independent characteristic.

### 3.6 Discussion

**Case study** We manually examined a number of samples, which can be successfully grouped by ADDML but not the baselines. For example, two aspect phrases "*photo quality*" and "*quality*" belong to group "*picture*" and "*build quality*", respectively. Most of the baselines incorrectly clustered them into the same group, while ADDML correctly grouped them. There are two main reasons: (1) the two aspect phrase themselves have similar semantics characteristic and share words, (2) reviewers commonly used similar words to express their opinion. ADDML can learn an exact vector representation that are context sensitive, while the baseline methods can not distinguish similar contexts. Figure 2 shows some example results of attention values.

	Purity						Entropy					
	DC	DVD	MP3	PHONE	avg	↑	DC	DVD	MP3	PHONE	avg	↑
AP	0.4196	0.4253	0.3600	0.4588	0.4159		2.1816	2.2074	2.3087	1.9946	2.1731	
<i>cnn + ml</i>	0.4673	0.4609	0.3003	0.4947	0.4308	3.6%	1.8984	1.9515	2.2908	1.6294	1.9433	10.6%
<i>atn + ml</i>	0.4605	0.4706	0.3070	0.5088	0.4367	5.0%	1.8477	1.8158	2.2662	1.5179	1.8619	14.3%
<i>cnn + mlp + ml</i>	0.5526	0.4118	0.3246	0.6140	0.4757	14.4%	1.8494	1.8980	2.1626	1.3336	1.8109	16.7%
<i>atn + mlp + ml(ADDML)</i>	<b>0.5658</b>	<b>0.5098</b>	<b>0.3684</b>	<b>0.6143</b>	<b>0.5146</b>	23.7%	<b>1.7119</b>	<b>1.8043</b>	<b>2.1274</b>	<b>1.3282</b>	<b>1.7429</b>	19.8%

Table 3: The result of different module combinations.

**Module Analysis** ADDML has three modules: attention-based semantic composition module (*atn*), MLP-based nonlinear transformation module (*mlp*) and metric learning (*ml*). For studying the contribution of each module, we introduce a general convolution neural network (*cnn*) as an alternative to *atn*. *cnn* is a state-of-the-art neural network method for modelling semantic representation of sentence (Kalchbrenner et al., 2014; Tang et al., 2015), which extracts N-gram features by convolution operations.

Table 3 reports the results of different module combinations. We use AP, which only uses aspect phrase vector for clustering, as a reference. The symbol  $\uparrow$  denotes average percentage improvement than AP in 4 domains. By considering context, *cnn + ml* and *atn + ml* achieved better results than AP, which only uses aspect phrase embeddings. After adding nonlinear transformation module, *cnn + mlp + ml* and *atn + mlp + ml* further improve the performance. Under the same condition, *atn* is superior to *cnn* for our task.

Naturally, aspect phrases in some domains, such as MP3, may have fixed meanings. As a result, an aspect phrase and its context have less correlation under the grouping task in these domains. Therefore, AP achieves a little better result than the other methods except for ADDML. Overall, *atn* solves the context representation and *mlp + ml* provides a better metric learning ability for our model.

**Similarity Threshold** Different similarity thresholds  $\eta$  results in different negative sample pairs, and have a certain impact on performance of our model. For obtaining a better threshold, we performed experiments on developing data with different similarity value. Figure 3 presents the result on DC dataset. The performance slowly decrease with the growth of threshold, which is in line with intuitively understanding. For obtaining enough negative samples, we chosen 0.3 as the similarity threshold.

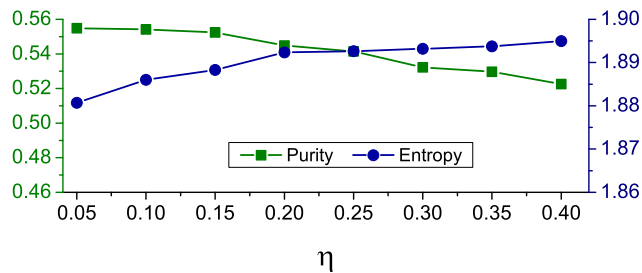


Figure 3: Influence of the similarity threshold  $\eta$

## 4 Related Work

Our work is related to aspect-level sentiment analysis, metric learning, and deep learning.

For **aspect-level sentiment analysis**, there are many methods on clustering aspect phrases. Some topic-model-based approaches jointly 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). 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 they did not perform well even considering pre-existing knowledge. Some other work focuses on grouping aspect phrases. Guo et al. (2009) grouped aspect phrases using multi-level LaSA, which exploits the



virtual context documents and semantic structure of aspect phrase. 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 better initialization for EM. Zhao et al. (2014) proposed a framework of Posterior Regularization to cluster aspect phrases, which formalizes sentiment distribution consistency as a soft constraint. This method requests special semi-structured reviews to estimate the sentiment distribution. In contrast to these methods, we provide a Siamese neural network to learning feature representation through distance supervising.

**Metric learning** algorithms have been successfully applied to address the problem of face verification (Ding et al., 2015; Yi et al., 2014; Cai et al., 2012; Hu et al., 2014). A common objective of these methods is to learn a better distance metric so that the distance between a positive pair is smaller than the distance between a negative pair. However, these methods not perform nonlinear transformation. Hu et al. (2014) employed a MLP-based nonlinear transformation, but its input is the given image descriptor, which can be directly concatenated to form feature vectors. In this paper, we adapt this method to the 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.

Our work is related to **word embedding** and **deep learning**. Prior research (Collobert and Weston, 2008; Mnih and Hinton, 2007; Mikolov et al., 2013; Tang et al., 2014; Ren et al., 2016b) presented different models to improve the performance of word embedding training, and our training is inspired by negative sampling. Deep learning methods (Kalchbrenner et al., 2014; Kim, 2014; Socher et al., 2013; Vo and Zhang, 2015; Zhang et al., 2015; Zhang et al., 2016; Ren et al., 2016a) have been applied to many tasks related to sentiment analysis. In this paper, we explore attention (Luong et al., 2015; Rush et al., 2015; Ling et al., 2015) with a MLP network to tackle the aspect phrase grouping problem.

## 5 Conclusion

We studied distance metric learning for aspect phrase grouping, exploring a novel deep neural network framework. By leveraging semantic relations between aspect phrase and their contexts, our approach give better performance to strong baselines which achieve the best results in standard benchmark. Our method can be applied to other NLP applications, such as short text clustering and sentence similarity measures.

## Acknowledgement

This work is supported by the National Natural Science Foundation of China (No. 61173062, 61373108, 61133012), the major program of the National Social Science Foundation of China (No. 11&ZD189), the key project of Natural Science Foundation of Hubei Province, China (No. 2012FFA088), the Educational Commission of Henan Province, China (No. 17A520050), the High Performance Computing Center of Computer School, Wuhan University and T2MOE201301 from Singapore Ministry of Education.

## References

- [Andrzejewski et al.2009] D. Andrzejewski, X. Zhu, and M. Craven. 2009. Incorporating domain knowledge into topic modeling via dirichlet forest priors. *Proc Int Conf Mach Learn*, 382(26):25–32.
- [Cai et al.2012] Xinyuan Cai, Chunheng Wang, Baihua Xiao, Xue Chen, and Ji Zhou. 2012. 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.
- [Chen et al.2013] Zhiyuan Chen, Arjun Mukherjee, Bing Liu, Meichun Hsu, Mal Castellanos, and Riddhiman Ghosh. 2013. Exploiting domain knowledge in aspect extraction. In *EMNLP*, EMNLP, pages 1655–1667. ACL.
- [Collobert and Weston2008] Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *ICML*, ICML, pages 160–167, New York, NY, USA.

- [Collobert et al.2011] Ronan Collobert, Jason Weston, L. E. On Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537.
- [Ding et al.2015] Shengyong Ding, Liang Lin, Guangrun Wang, and Hongyang Chao. 2015. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*.
- [Faruqui et al.2015] Manaal Faruqui, Jesse Dodge, Sujay K. Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of NAACL*, NAACL.
- [Glorot and Yoshua2010] Xavier Glorot and Bengio Yoshua. 2010. 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.
- [Guo et al.2009] Honglei Guo, Huijia Zhu, Zhili Guo, XiaoXun Zhang, and Zhong Su. 2009. Product feature categorization with multilevel latent semantic association. In *CIKM*, CIKM, pages 1087–1096. ACM.
- [Hu and Liu2004] Mingqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. *KDD*, pages 168–177. ACM.
- [Hu et al.2014] Junlin Hu, Jiwen Lu, and Yap-Peng Tan. 2014. Discriminative deep metric learning for face verification in the wild. In *cvpr*, cvpr, pages 1875 – 1882. IEEE.
- [Jin et al.2009] Wei Jin, Hung Hay Ho, and Rohini K. Srihari. 2009. Opinionminer: A novel machine learning system for web opinion mining and extraction. In *KDD '09*, KDD '09, pages 1195–1204, New York, NY, USA. ACM.
- [Jindal and Liu2008] Nitin Jindal and Bing Liu. 2008. Opinion spam and analysis. In *WSDM*, WSDM, pages 219–230. ACM.
- [Jo and Oh2011] Yohan Jo and Alice H. Oh. 2011. Aspect and sentiment unification model for online review analysis. In *Proceedings of WSDM*, WSDM '11, pages 815–824, New York, NY, USA. ACM.
- [Kalchbrenner et al.2014] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. In *ACL*, ACL.
- [Kim and Hovy2006] Soo-Min Kim and Eduard Hovy. 2006. 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. Association for Computational Linguistics.
- [Kim2014] Yoon Kim. 2014. 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.
- [Kobayashi et al.2007] Nozomi Kobayashi, Kentaro Inui, and Yuji Matsumoto. 2007. Extracting aspect-evaluation and aspect-of relations in opinion mining. pages 1065–1074, Prague, Czech Republic. Association for Computational Linguistics.
- [Lin and He2009] Chenghua Lin and Yulan He. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceedings of CIKM*, CIKM, pages 375–384, Hong Kong, China. 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. 2015. Not all contexts are created equal: Better word representations with variable attention. In *Proceedings of EMNLP*, EMNLP, pages 1367–1372, Lisbon, Portugal.
- [Lu et al.2011] Bin Lu, M. Ott, C. Cardie, and B. K. Tsou. 2011. Multi-aspect sentiment analysis with topic models. In *Proc. ICDMW*, Proc. ICDMW, pages 81 – 88. IEEE.
- [Luong et al.2015] Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, EMNLP, pages 1412–1421, Lisbon, Portugal.
- [Mignon and Jurie2012] A. Mignon and F. Jurie. 2012. Pcca: A new approach for distance learning from sparse pairwise constraints. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2666–2672.

- [Mikolov et al.2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proc. NIPS*, Proc. NIPS, pages 3111–3119. Curran Associates, Inc.
- [Mnih and Hinton2007] Andriy Mnih and Geoffrey Hinton. 2007. Three new graphical models for statistical language modelling. In *ICML*, ICML, pages 641–648.
- [Moghaddam and Ester2012] Samaneh Moghaddam and Martin Ester. 2012. On the design of lda models for aspect-based opinion mining. In *Proceedings of CIKM*, CIKM, pages 803–812. ACM. 2396863.
- [Pang and Lee2008] Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.
- [Pedersen2010] Ted Pedersen. 2010. Information content measures of semantic similarity perform better without sense-tagged text. In *HLT '10*, HLT '10, pages 329–332, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [Pennington et al.2014] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proc. EMNLP*, Proc. EMNLP, pages 1532–1543. Association for Computational Linguistics.
- [Ren et al.2016a] Yafeng Ren, Yue Zhang, Meishan Zhang, and Donghong Ji. 2016a. Context-sensitive twitter sentiment classification using neural network.
- [Ren et al.2016b] Yafeng Ren, Yue Zhang, Meishan Zhang, and Donghong Ji. 2016b. Improving twitter sentiment classification using topic-enriched multi-prototype word embeddings.
- [Rush et al.2015] Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, EMNLP, pages 379–389, Lisbon, Portugal.
- [Socher et al.2013] Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, page 1642. Citeseer.
- [Tang et al.2014] Duyu Tang, Furu Wei, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for twitter sentiment classification. In *ACL*, ACL.
- [Tang et al.2015] Duyu Tang, Bing Qin, and Ting Liu. 2015. Learning semantic representations of users and products for document level sentiment classification. In *Proceedings of ACL*, pages 1014–1023, Beijing, China.
- [Vo and Zhang2015] Duy-Tin Vo and Yue Zhang. 2015. Deep learning for event-driven stock prediction. In *Proceedings of IJCAI*, IJCAI, BueNos Aires, Argentina.
- [Xiong and Ji2015] Shufeng Xiong and Donghong Ji. 2015. Exploiting capacity-constrained k-means clustering for aspect-phrase grouping. In Songmao Zhang, Martin Wirsing, and Zili Zhang, editors, *Knowledge Science, Engineering and Management*, volume 9403 of *Knowledge Science, Engineering and Management*, pages 370–381. Springer International Publishing.
- [Xu et al.2015] Jiaming Xu, Peng Wang, Guanhua Tian, Bo Xu, Jun Zhao, Fangyuan Wang, and Hongwei Hao. 2015. Short text clustering via convolutional neural networks. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 62–69, Denver, Colorado.
- [Yi et al.2014] Dong Yi, Zhen Lei, and Stan Z. Li. 2014. Deep metric learning for practical person re-identification. *arXiv preprint arXiv:1407.4979*.
- [Yu and Dredze2014] Mo Yu and Mark Dredze. 2014. Improving lexical embeddings with semantic knowledge. In *Proceedings of ACL*, pages 545–550, Baltimore, Maryland.
- [Zhai et al.2010] Zhongwu Zhai, Bing Liu, Hua Xu, and Peifa Jia. 2010. Grouping product features using semi-supervised learning with soft-constraints. In *Proc. COLING*, Proc. COLING, pages 1272–1280.
- [Zhai et al.2011a] Zhongwu Zhai, Bing Liu, Hua Xu, and Peifa Jia. 2011a. Clustering product features for opinion mining. In *Proc. WSDM*, Proc. WSDM, pages 347–354. ACM. 1935884.
- [Zhai et al.2011b] Zhongwu Zhai, Bing Liu, Hua Xu, and Peifa Jia. 2011b. Constrained lda for grouping product features in opinion mining. In *Proc. PAKDD*, Proc. PAKDD, pages 448–459.

- [Zhang et al.2015] Meishan Zhang, Yue Zhang, and Duy Tin Vo. 2015. Neural networks for open domain targeted sentiment. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 612–621, Lisbon, Portugal.
- [Zhang et al.2016] Meishan Zhang, Yue Zhang, and Duy-Tin Vo. 2016. Gated neural networks for targeted sentiment analysis. In *AAAI, AAAI*.
- [Zhao et al.2010] Wayne Xin Zhao, Jing Jiang, Hongfei Yan, and Xiaoming Li. 2010. Jointly modeling aspects and opinions with a maxent-lda hybrid. In *Proceedings of EMNLP, EMNLP*, pages 56–65. Association for Computational Linguistics. 1870664.
- [Zhao et al.2014] Li Zhao, Minlie Huang, Haiqiang Chen, Junjun Cheng, and Xiaoyan Zhu. 2014. Clustering aspect-related phrases by leveraging sentiment distribution consistency. In *Proc. EMNLP, Proc. EMNLP*, pages 1614–1623. Association for Computational Linguistics.