Hyperbolic Disentangled Representation for Fine-Grained Aspect Extraction

Chang-You Tai\textsuperscript{1}, Ming-Yao Li\textsuperscript{1}, Lun-Wei Ku\textsuperscript{1}

\textsuperscript{1}Academia Sinica, Taipei, Taiwan
johnnyjana730@gmail.com, xuaul@gmail.com, lwku@iis.sinica.edu.tw

Abstract

Automatic identification of salient aspects from user reviews is especially useful for opinion analysis. There has been significant progress in utilizing weakly supervised approaches, which require only a small set of seed words for training aspect classifiers. However, there is always room for improvement. First, no weakly supervised approaches fully utilize latent hierarchies between words. Second, each seed word’s representation should have different latent semantics and be distinct when it represents a different aspect. In this paper we propose HDAE, a hyperbolic disentangled aspect extractor in which a hyperbolic aspect classifier captures words’ latent hierarchies, and an aspect-disentangled representation models the distinct latent semantics of each seed word. Compared to previous baselines, HDAE achieves average F1 performance gains of 18.2\% and 24.1\% on Amazon product review and restaurant review datasets, respectively. In addition, the embedding visualization experience demonstrates that HDAE is a more effective approach to leveraging seed words. An ablation study and a case study further attest the effectiveness of the proposed components.

Introduction

Researchers have begun to focus on aspect extraction, the automatic detection of fine-grained segments with predefined aspects \cite{Hu and Liu 2004, Liu 2012, Pontiki et al. 2016}, due to its potential for downstream tasks. For example, aspect extraction benefits users and customers when searching through review segments for aspects of interest on the Internet. Aspect extraction is also crucial for document summarization \cite{Angelidis and Lapata 2018}, recommendation justification \cite{Ni, Li, and McAuley 2019}, and review-based recommendation \cite{Chin et al. 2018}.

Aspect extraction research can be divided into supervised approaches, unsupervised approaches, and weakly supervised approaches\cite{Chang-You Tai et al.} Among these, many studies have been conducted on weakly supervised approaches \cite{Karamanolakis et al. 2019, Angelidis and Lapata 2018, Zhuang et al. 2020} since they allow the model to be trained without substantial human-labeled data. For example, \cite{Angelidis and Lapata 2018} initialize fine-grained aspect representations using only a small number of descriptive keywords, or seed words, to identify highly salient opinions in review segments. Also, \cite{Karamanolakis et al. 2019} propose a student-teacher framework that more effectively leverages seed words by using a bag-of-words classifier teacher.

However, there is room for improvement in such seed word based methods. First, they neglect to consider the latent hierarchies between words, and it is assumed that capturing latent hierarchies between words will further improve seed word based methods on aspect inference, for instance by better identifying and organizing seed words and their hypernym pairs \cite{Huang et al. 2020, Lopez, Heinzerling, and Strube 2019}. For example, as shown in Fig. 1(a), the general seed word color near the top can be used to find the more specific words blue or green in the middle, after which even more specific words can be found such as ultramarine or azure celeste. If seed words or their hypernym pairs exist in one review segment, the model can infer that it is of the corresponding aspect.

To allow the model to fully capture latent hierarchies between words, we introduce hyperbolic space \cite{Nickel and Kiela 2017, Murty et al. 2018, Xu and Barbosa 2018, Lopez, Heinzerling, and Strube 2019, Lopez and Strube 2020}. Compared to Euclidean space, hyperbolic space effectively encodes hierarchical structure information \cite{Nickel and Kiela 2017}, the latent hierarchies between words in this paper. In particular, when embedding tree-like structures, compared to the volume in Euclidean space, which leads to high distortion embeddings \cite{Sa et al. 2018, Sarkar 2011}, volume in hyperbolic space grows exponentially and can embed trees with arbitrarily low distortion \cite{Sarkar 2011, Nickel and Kiela 2017}. By virtue of such a hierarchy, a seed word based model

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure1}
\caption{a) Seed word color and its hypernym pairs. b) An illustration of latent semantics under seed word picture. For example, in the TV domain’s image aspect, pixel of picture $s_{i,j}^{d_1}$ and screen picture $s_{i,j}^{d_2}$ exist, whereas in the boot domain’s look aspect, cute picture $s_{i,j}^{d_1}$ and attractive picture $s_{i,j}^{d_2}$ exist.}
\end{figure}
can better identify and utilize seed words and their hypernym words and thus achieve better aspect inference in hyperbolic space.

Second, existing seed word-based approaches model each seed word representation in a uniform manner while neglecting the fact that each seed word should have different latent semantics when conducting aspect extraction. For example, for the Amazon product review dataset (Angelidis and Lapata 2018), in the TV domain’s picture aspect, the latent semantics under the seed word picture can be pixel of the picture, screen picture, or HD picture, as shown in Fig. 1(b). It is essential to select the most relevant latent semantics of the seed word when using the seed word picture to infer review aspects of segments. Furthermore, as shown in Fig. 1(b), the latent semantics of the seed word should be different in different aspects: this is also neglected by the current uniform representation. Such a uniform approach to modeling seed words tends to result in sub-optimal representations.

Thus, we propose HDAE, a hyperbolic disentangled aspect extractor which captures words’ latent hierarchies and disentangles the latent semantics of each seed word. First, we propose a hyperbolic aspect classifier, using a hyperbolic distance function to calculate the relationship between the segment vector and the aspect representation generated from the seed word. Second, we introduce an aspect disentanglement module to model each seed word’s latent semantics and then generate an aspect-refined representation of each review segment by selecting the most relevant latent semantics. In addition, we propose aspect-aware regularization to model each latent semantic meaning under its aspect scope while encouraging the independence of different latent semantic meanings. We conduct experiments on two datasets, demonstrating that HDAE achieves better aspect inference, which is further substantiated by embedding visualizations. We also provide two case studies to investigate HDAE’s aspect inference ability compared with baselines without fully capturing words’ latent hierarchies and the interpretability of the seed words’ disentangled latent semantics.

We summarize our contributions: first, we propose a novel hyperbolic disentangled aspect extractor. To the best of our knowledge, this is the first work to investigate how to leverage hyperbolic components and disentangled representations for weakly supervised approaches to aspect extraction. Second, we propose a hyperbolic aspect classifier which captures word’s latent hierarchies and generates associations between the review segment and aspects of interest. Third, we introduce the aspect disentanglement module and aspect-aware latent semantic regularization to model the latent semantic meaning of each seed word. Experiments and a case study demonstrate the effect of the proposed methods for aspect extraction.

Related Work

Aspect Extraction In addition to weakly supervised approaches, there are also supervised approaches and unsupervised approaches. Supervised neural networks achieve better performance than traditional rule-based approaches by viewing aspect extraction as a sequence labeling problem which can be tackled with hidden Markov models (Jin, Ho, and Srihari 2009), conditional random fields (Yang and Cardie 2012), or recurrent neural networks (Wang et al. 2016; Liu, Joty, and Meng 2015). However, supervised approaches require large amounts of labeled data for training. Unsupervised approaches, in contrast, do not use annotated data. Early examples are latent Dirichlet allocation (LDA)-based methods (Chen, Mukherjee, and Liu 2014; García-Pablos, Cuadros, and Rigau 2018; Shi et al. 2018; Sun et al. 2018), in the TV domain, as shown in Fig. 1(b). It is essential to select the most relevant latent semantics of the seed word when using the seed word picture to infer review aspects of segments. Furthermore, as shown in Fig. 1(b), the latent semantics of the seed word should be different in different aspects: this is also neglected by the current uniform representation. Such a uniform approach to modeling seed words tends to result in sub-optimal representations.

Hyperbolic representations have been used to model complex networks (Krioukov et al. 2010; Nickel and Kiela 2018) and have proven more suitable than Euclidean space in representing hierarchical data (Sala et al. 2018; Nickel and Kiela 2017). For example, Lopez and Strube (2020) introduce hyperbolic representations to capture latent hierarchies arising from the class distribution for multi-class multi-label classification. Aly et al. (2019) use Poincaré embeddings to improve existing methods for domain-specific taxonomy induction. Le et al. (2019) propose utilizing hyperbolic representations to infer missing hypernymy relations. Sun et al. (2021) show that points in hyperbolic space can be more concentrated while maintaining the desired separation and revealing nuanced differences. To our knowledge, this is the first work to apply hyperbolic representations to weakly supervised approaches for aspect extraction.

Disentangled representations improve model performance by identifying and disentangling latent explanatory factors in the observed data (Yoshua Bengio and Vincent 2012) and have shown their success in the NLP domain (Shen et al. 2017; Zhao et al. 2018; Chen et al. 2019; Hu et al. 2017). For instance, Hu et al. (2017) propose disentangled representations with designated semantic structure, which generates sentences with dynamically specified attributes. Hou et al. (2021) derive disentangled representations which separate the distinct and informative factors of variations to improve content-based detection. Disentangled representation has been successively applied to the recommendation (Ma et al. 2019) and computer vision (Liu et al. 2020; Dupont 2018) domains. For example, Wang et al. (2020) model diverse relationships and disentangle user intents to achieve better-performing representations. To our knowledge, this is the first work to apply disentangled representations to weakly supervised approaches for aspect extraction.

2The codes is at https://github.com/johnnyjana730/HDAE/
Preliminaries

Problem formulation The goal of aspect extraction is to predict an aspect category \( a_i \in A_C = \{a_j\}_{j=1}^K \) given a review segment (e.g., sentence, clause) \( x^a = \{x_1, x_2, \ldots, x_T\} \) from a specific domain \( d_C \) (e.g., laptop bags, TVs), where the review segments are created by splitting each review in the corpus; \( x_i \) is the word index in the segment; \( a_i \) is an aspect and \( A_C \) refers to the aspect set pertaining to domain \( d_C \); \( K \) is the number of total aspects and \( A_C \) refers to the segment’s length. For every aspect \( a_i \in A_C \), a small number of seed words \( \{s_{i,1}, s_{i,2}, \ldots, s_{i,N}\} \) are provided during training. The classifier predicts \( K \) aspect probabilities \( p^a_s = \{p^a_s, \ldots, p^K_a\} \) given a test review segment \( x^a \) and the seed words.

Hyperbolic Geometry We introduce two hyperbolic models, the Poincaré ball model and the Klein model.

The Poincaré ball model is defined as a Riemannian manifold \( P^n = (\beta, g^P) \), where \( \beta^n = \{x \in \mathbb{R}^n : ||x|| < 1\} \) is an open unit ball, with the metric tensor \( g^P = \lambda^2 g^E \), where \( \lambda = 1 - ||x||^2 \) is the Euclidean metric tensor. The distance on the manifold is defined as
\[
d_P(x,y) = \arccosh \left( 1 + 2 \frac{||x-y||^2}{1-||x||^2(1-||y||^2)} \right). \tag{1}
\]

The Klein model is given by \( K^n = \{x \in \mathbb{R}^n : ||x|| < 1\} \) and is often used for aggregation since the Einstein midpoint [Gulcehre et al. 2018] can be easily computed in the Klein model. Formally, a point in the Klein model can be obtained from Poincaré coordinates by
\[
P^n \rightarrow K^n : \pi_P \rightarrow K(x_P) = \frac{2x_P}{1 + ||x_P||^2} \tag{2}
\]
and the backward transition formulas
\[
K^n \rightarrow P^n : \pi_K \rightarrow P(x_K) = \frac{x_K}{1 + \sqrt{1 - ||x_K||^2}}. \tag{3}
\]

For the Poincaré ball model, the exponential map, from tangent space to hyperboloid manifold, \( \exp_P : T_P \rightarrow P, \) and the logarithmic map, from hyperboloid manifold to tangent space, \( \log_P : P \rightarrow T_P \), can be found in [Liu, Nickel, and Kiela 2019]. For simplicity, we denote \( d_P^{exp} \) as the hyperbolic distance of the tangent space vector after applying the exponential map:
\[
d_P^{exp}(x,y) = d_P(\exp_P(x), \exp_P(0)). \tag{4}
\]

Methodology

Euclidean Aspect Extractor

Our work builds on the seed word based model developed by [Angelidis and Lapata]. We describe the method, including segment representation generation and the aspect classifier.

Segment Representation For each review segment \( x^a = \{x_1, x_2, \ldots, x_T\} \), the segment representation \( v_s \) is generated by a weighted sum of an individual word:
\[
v_s = \sum_{i=1}^n c_i x_{i,s} \tag{5}
\]
\[
c_i = \frac{\exp(u_i)}{\sum_{j=1}^n \exp(u_j)}; u_i = v_{x_i}^\top \cdot M \cdot v_s^\top \tag{6}
\]
\[\text{For more details; see Robbin and Salamon [2011].}\]

where \( v_{x_i} \) is the vector of the \( i \)-th word \( x_i; v_s^\top \) is average of the segment’s word vector; and \( M \in \mathbb{R}^{d \times d} \) denotes the attention matrix.

Euclidean Aspect Classifier To predict a probability distribution over \( K \) aspects, the vector \( v_s \) is fed to a hidden classification layer followed by the softmax function:
\[
p^s_a = \text{softmax}(Wv_s + b), \tag{7}
\]
where \( W \) and \( b \) are trainable parameters. To focus on the aspect of interest, for each aspect \( a_i \), which has seed words \( \{s_{i,1}, s_{i,2}, \ldots, s_{i,N}\} \), the model generates the aspect vector \( a_i \) using the labeled aspect seed words:
\[
a_i = \sum_{j}^N z_{i,j} s_{i,j}; A = [a_1^\top; \ldots; a_K^\top], \tag{8}
\]
where \( A \in \mathbb{R}^{K \times d} \) denotes the aspect matrix; and \( s_{i,j} \) denotes the \( j\)-th seed word representation; the weight vectors \( z_{i,j} \) are determined by the method mentioned in [Angelidis and Lapata 2018]; and \( N \) is the number of seed words. Then, the segment reconstructed vector \( r_s \) is generated based on the aspect vector:
\[
r_s = A^\top p^s_{asp}. \tag{9}
\]

To optimize the performance, the model is trained by reconstruction loss, which maximizes the distance between inner product \( r_s \cdot v_s \) and \( r_s \cdot v_{n_i} \), where \( v_{n_i} \) is the vector of a randomly sampled negative segment.
\[
J_r(\theta) = \sum_{x^a \in C} \sum_{i=1}^k \max(0, 1 - r_s \cdot v_s + r_s \cdot v_{n_i}), \tag{10}
\]

Hyperbolic Disentangled Aspect Extractor

Here, we present HDAE, a hyperbolic aspect classifier with an aspect disentanglement module proposed to model multiple latent semantic meanings for each seed word according to its aspect category.

Hyperbolic Aspect Classifier To infer the review segment’s aspect probability \( p^s_a \) in hyperbolic space, we follow [Balažević, Allen, and Hospedales 2019] in using the hyperbolic distance function and biases to calculate the relationship between segment vector \( v_s \) and aspect representation \( a_i \) as
\[
p^s_a = -d_P^{exp}(v_s, a_i)^2 + b_v + b_a. \tag{11}
\]

Then, to generate the reconstructed embedding \( r_s \), the Einstein midpoint is used to aggregate hyperbolic aspect weights, with a simple form in the Klein disk model:
\[
r_s = \log_0(\pi_{K \rightarrow P}(\sum_{a_i \in A_C} \sum_{j=1}^K k_j \gamma(a^K_j) a^K_j)) \tag{12}
\]
\[
k_i = \exp(\beta p^s_a - c), \tag{13}
\]
where \( a^K_i = \pi_{P \rightarrow K}(a^P_i); a^P_i \) denotes the Poincaré aspect embedding; \( a^P = \exp_0(a_i) \); \( \beta \) and \( c \) are set parameters; and Lorentz factors \( \gamma(t) = \frac{1}{(1-||t||^2)^{1/2}}. \)
Aspect Disentanglement Module  To generate multiple latent semantic meanings for each seed word, we propose a disentangled semantic representation. Then, we present aspect-aware regularization, which models latent semantic vectors for each seed word, after which we discuss refined seed word representation.

Disentangled Semantic Representation For aspect \( a_i \), we devise a representation function to output a disentangled semantic vector \( s_{i,j}^d \) for the \( j \)-th seed word \( s_{i,j} \), which is composed of \( I \) independent components:

\[
s_{i,j}^d = (s_{i,j}^{d_1}, s_{i,j}^{d_2}, s_{i,j}^{d_3}, \ldots, s_{i,j}^{d_I}),
\]

where disentangled semantic vector \( s_{i,j}^{d_k} \) is generated by adding a standard Gaussian random variable to the original seed word representation \( s_{i,j} \).

Aspect-Aware Regularization  This models the latent semantic representation of each seed word according to its aspect category and has three objectives, as shown in Fig. 2: (a) seed word dependence, (b) latent semantic independence, and (c) aspect scope confinement, which are controlled by latent semantic modeling distances \( d_1, d_2, \) and \( d_3 \).

Seed Word Dependence  The interdependence between seed word pairs sheds light on the modeling of the seed word’s latent semantics within the scope of its aspect. For example, for seed word design in the boot domain’s look aspect, the latent semantic meaning, which facilitates fine-grained aspect inference, can be color design, design style, cute design, and attractive design. The desired latent semantic meaning can be modeled by narrowing the gap between either the latent semantic meaning of design and the latent semantic meanings of other seed words, such as color, style, cute, and attractive in the same look aspect. Likewise, in the TV domain’s service aspect, latent semantic meanings shipping service, replacement service, and delivery service can be generated by minimizing the distance between either the latent semantic meaning of service and that of shipping, replacement, and delivery, which are seed words in the same aspect.

To model the interdependence of seed word pairs, we use the hyperbolic distance function \( d_P(\cdot) \) to achieve fine-grained relationship modeling, since hyperbolic space offers the ability to not only preserve hierarchical (tree-like) information (Nickel and Kiela 2017; Zhang and Gao 2020; Gulcehre et al. 2018; Chami et al. 2019) but also nuanced differences (to better group them) (Sun et al. 2021; Tai et al. 2021) and outperforms Euclidean counterparts in various kinds of data (Zhang and Gao 2020; Gulcehre et al. 2018; Chami et al. 2019) 2020; Sun et al. 2021; Tai et al. 2021). Thus, it is assumed that with more space (hyperbolic space) to organize points, the model can divide disentangled representations and better group them. Given seed word pairs such as design \( s_{i,j} \) and color \( s_{i,j'} \) in the specific aspect, we require at least one latent semantic pair distance to be close enough:

\[
\begin{align*}
\text{sim}(s_{i,j}, s_{i,j'}) &= \arg\min\{d^\text{exp}_P(s_{i,j}, s_{i,j'}) | s_{i,j}, s_{i,j'} \in s_{i,j}^d \} \\
J_{d_1}(\theta) &= \sum_{a_i \in A_C} \sum_{j=1}^N \sum_{j' = j+1}^N \max(0, (0, d_2 - d^\text{exp}_P(s_{i,j}, s_{i,j'})))
\end{align*}
\]

where \( \text{sim} \) outputs the minimal distance from all possible seed word latent semantic meaning pairs; \( d_1 \) is the inter seed word alignment distance, which maintains two latent semantic meanings within a certain distance. Intuitively, for different aspect word pairs, the alignment score should be different, as in Wang et al. (2020). For example, in the boot domain’s look aspect, the seed word dependence between design and color should be more significant than design and going. We leave this to future work.

Latent Semantic Independence  Latent semantic meanings should be distinct from each other. Independent latent semantic meanings reduce redundancy and confusion in aspect inference. To achieve this, we maintain the distance between the seed word’s latent semantic meanings.

\[
J_{d_2}(\theta) = \sum_{a_i \in A_C} \sum_{j=1}^N \sum_{k=1}^1 \max(0, (0, d_2 - d^\text{exp}_P(s_{i,j}, s_{i,j'})))
\]

where \( d_2 \) is the latent semantic distance.

Aspect Scope Confinement  For each seed word, all latent semantic meanings should be limited in terms of aspect scope. For example, in the boot domain’s color aspect, all latent semantic meanings of seed word style should refer to color’s style. However, in the look aspect, all latent semantic meanings of the same seed word style should refer to outlook style. To thus limit all latent meanings of a seed word to its aspect scope, we introduce another regularization:

\[
J_{d_3}(\theta) = \sum_{a_i \in A_C} \sum_{j=1}^N \max(0, d^\text{exp}_P(s_{i,j}, a_i) - d_3),
\]

where \( d_3 \) is the aspect scope confinement distance and \( a_i \) is the aspect representation from Eq. 5. Note compared to seed word dependence and Eq. 16, which focuses on dependence between seed word pairs, Eq. 18 ensures all latent semantic meanings are modeled within the specific aspect scope.
Refined Seed Word Representation This constructs refined seed representations based on its latent semantics. For each seed word, the latent semantics should be independent from each other; only one latent semantic meaning should be used to find the aspect relevant content. For example, for the boot domain’s look aspect, possible latent semantics of seed word style include cute style, casual style, or attractive style; as we can see these latent semantics are of different meanings, and combining them together may lead to a sub-optimal seed word representation. Thus, we introduce when predicting its aspect distribution. Therefore, we introduce 

$$s_{i,j} = \sum_{k=1}^{G} g_k s_{i,j}^k, \quad g_k = \frac{c_k}{\sum_{k'} c_{k'}}.$$  

(19)

c_k = \exp\left(-d_p(v_{s,k}, s_{i,j}^k) / \tau \right).  

(20)

where \( \tau \), the temperature parameter, controls the extent to which the output becomes a one-hot vector. With the refined seed word representation \( s_{i,j} \) according to each segment \( v_{s} \), the aspect representation can be generated by Eq. 8.

Algorithm 1: HDAE Learning

| Input: | review segments \( S = \{s \mid s \in C \} \), aspect seed words |
|---|---|
| 1 | Initialize HDAE parameter with pre-trained word vector |
| 2 | foreach epoch do |
| 3 | for \( x^k \in S \) do |
| 4 | Generate segment embedding \( v_{s} \). (Eq 5) |
| 5 | for \( i \leftarrow 1 \) to \( K \) do |
| 6 | Generate refined aspect seed word vector \( s_{i,j}^k \). (Eq 19) |
| 7 | Calculate aspect embedding \( a_{i} \). (Eq 15) |
| 8 | Generate aspect probability \( p_i^{d_i} \). (Eq 11) |
| 9 | Generate reconstructed embedding \( r_{s} \). (Eq 12) |
| 10 | Calculate objective \( J \). (Eq 21) |
| 11 | Update parameters by Adam optimizer |

Learning Algorithm

The formal description of the above aspect inference process is presented in Algorithm 1. To train HDAE, we rely on the previously introduced reconstruction loss \( J_r \) (Eq. 10). Since the reconstruction objective only provides a weak training signal (Angelidis and Lapata 2018), the distillation objective \( J_d \) from the teacher (Karamanolakis et al. 2019) is used to provide an additional training signal. Also, the disentangled modeling objectives \( J_{d_1} \), \( J_{d_2} \), and \( J_{d_3} \) are used to model each latent semantic meaning according to its aspect category. Thus, the overall objective is

$$J(\theta) = J_r(\theta) + \lambda J_d(\theta) + J_{d_1}(\theta) + J_{d_2}(\theta) + J_{d_3}(\theta).$$  

(21)

The \( \lambda \) controls the influence of the distillation objective loss.

Experiments and Results

Datasets We used Amazon product reviews from the OPOSUM dataset (Angelidis and Lapata 2018) and restaurant reviews from the SemEval-2016 Aspect-based Sentiment Analysis task (Pontiki et al. 2016). The Amazon product reviews cover six domains, ranging from laptop bags (Bags), Bluetooth headsets (BT), boots, keyboards (KBs), and televisions (TVs) to vacuums (VCs). The restaurant reviews dataset covers six languages: English (En), Spanish (Sp), French (Fr), Russian (Ru), Dutch (Du), and Turkish (Tur). During training, seed words are provided but not segment aspect labels. Details are provided in the appendix.

Baseline LDA- Anchors (Lund et al. 2017), an interactive topic model which utilizes seed words as “anchors” to identify the segment aspect. ABAE (He and Chua 2017), an unsupervised method which adopts reconstruction loss to make the reconstructed embedding similar to a segment vector. This requires a manual mapping between the model-inferred aspect and gold-standard aspects. SSCL (Shi et al. 2020), an unsupervised method that uses a constraint learning algorithm and knowledge distillation for aspect inference. For manual mapping, the high-resolution selective mapping (HRSMap) is used. MATE* (Angelidis and Lapata 2018), a seed-based weakly supervised method which generates pre-defined aspect representations by seed word vector. This can be trained by an extra multitask training objective (MT) Ts.* (Karamanolakis et al. 2019), a seed-based weakly supervised method which adopts a teacher-student iterative co-training framework, where the teacher (TS-Teacher) is a bag-of-words classifier based on seed words and the student uses the attention-weighted average of word2vec embeddings (TS-ATT). Gold*., supervised models trained using ground truth aspect labels, only available for restaurant reviews, and not directly comparable with other weakly supervised baselines (Karamanolakis et al. 2019).

Note that for SSCL and TS, the BERT model also can be used as the encoder (SSCL-BT, TS-BT). The results of the compared models are obtained from the corresponding published papers. We also report our re-implemented version of SSCL-BT*. We do not provide the ABAE and SSCL results for restaurant reviews for non-English datasets, since this requires domain knowledge for manual aspect mapping.

Implementation Details For HDAE and other models, detailed hyper-parameter settings are given in the appendix.

Experimental Results

Overall Inference Performance Tables 1 and 2 show the results for aspect extraction on both datasets. We observe that HDAE achieves superior performance. For example, in Amazon product reviews, compared to TS-W2V, HDAE yields F1 performance gains of 16.0%, 8.1%, 31.9%, 24.7%, 11.3%, and 17.4% on Bags, KBs, Boots, BT, TVs, and VCs, respectively; similar trends are observed in the restaurant review dataset. Moreover, the reduction in the parameter size.

*MT cannot be applied in restaurant reviews since it requires datasets from different domains but the same language.

*We report ABAE and SSCL results for EN restaurant reviews in the appendix in our arxiv version.
Table 1: Micro-averaged F1 for 9-class EDU-level aspect detection in product reviews

| Model       | Bags  | KBs   | Boots | B/T  | TVs  | VCs  |
|-------------|-------|-------|-------|------|------|------|
| LDA-Anchors | 33.5  | 34.7  | 31.7  | 38.4 | 29.8 | 30.1 |
| ABAE        | 38.1  | 38.6  | 35.2  | 37.6 | 39.5 | 38.1 |
| SSCL        | 61.0  | 60.6  | 57.3  | 65.2 | 64.6 | 57.2 |
| SSCL-BT     | 65.5  | 62.3  | 60.4  | 69.5 | 67.0 | 61.0 |
| SSCL-BT*    | 56.5  | 61.7  | 41.5  | 51.4 | 58.2 | 52.4 |
| MATE        | 46.2  | 43.5  | 45.6  | 52.2 | 48.8 | 42.3 |
| MATE-MT     | 48.6  | 45.3  | 46.4  | 54.5 | 51.8 | 47.7 |
| TS-Teacher  | 59.3  | 58.2  | 50.6  | 63.3 | 61.0 | 58.4 |
| TS-ATT      | 58.7  | 57.0  | 52.6  | 67.6 | 63.2 | 58.8 |
| TS-BT       | 59.1  | 59.0  | 53.9  | 65.8 | 66.1 | 61.0 |
| HDAE        | 68.8  | 72.2  | 64.0  | 72.0 | 71.2 | 66.9 |

Table 2: Micro-averaged F1 for 12-class sentence-level aspect detection in restaurant reviews

| Model        | En | Sp | Fr | Ru | Du | Tur |
|--------------|----|----|----|----|----|-----|
| W2V-Gold     | 58.8| 50.4| 50.4| 69.3| 51.4| 55.7|
| BERT-Gold    | 63.1| 51.6| 50.6| 64.6| 53.5| 55.3|
| HDAE-Gold    | 70.5| 72.5| 65.4| 67.9| 73.8| 65.4|
| LDA-Anchors  | 28.5| 17.7| 13.1| 14.8| 25.9| 27.7|
| MATE         | 41.0| 24.9| 25.8| 18.4| 36.1| 39.0|
| MATE-UW      | 40.3| 18.3| 27.8| 21.8| 31.5| 25.2|
| TS-Teacher   | 44.9| 41.8| 34.1| 54.4| 40.7| 30.2|
| TS-ATT       | 47.8| 41.7| 32.4| 59.0| 42.1| 42.3|
| TS-BT        | 51.8| 42.0| 39.2| 58.0| 43.0| 45.0|
| HDAE         | 57.9| 65.7| 48.6| 62.9| 57.2| 50.8|

Table 3: HDAE ablation study. The $\lambda$ is the ratio of distillation objective loss. When $\lambda = 0$, the distillation objective $J_d$ is not used.

| Ablation   | Bag   | KBs   | B/T   | Boots | TV   | VCs  |
|------------|-------|-------|-------|-------|------|------|
| HDAE       | 68.8  | 72.2  | 72.0  | 64.0  | 71.2 | 66.9 |
| HDAE ($\lambda = 0$) | 67.3 | 65.6 | 70.1 | 60.5 | 54.1 | 59.1 |
| MATE       | 46.2  | 43.5  | 52.2  | 45.6  | 48.8 | 42.3 |

The parameter sizes of HDAE and TS-BT are 2.5M and 109.5M.

---

We also observe the weakly unsupervised approaches MATE* and TS-* significantly outperform the unsupervised approaches LDA-Anchors and ABAE, suggesting the effectiveness of seed words. Note our reproduced SSCL-BT* does not consistently outperform MATE, perhaps because SSCL-BT relies heavily on the quality of initial k-means centroids since poorly initialized centroids may cause model-inferred aspects after training to lack good coverage for gold-standard aspects, and thus make manual mapping more difficult.

To verify the effectiveness of the proposed components, we conducted an ablation study for HDAE, as shown in Table 3. After removing the hyperbolic aspect classifier (3) and aspect disentanglement module (4), we observe drops in performance, indicating the effect of the proposed components. Note that (4), which only contains the hyperbolic aspect classifier, out-
Table 4: Comparison of predictions on sample Product review segments between HDAE, MATE, and MATE.*. For each review segment, the ground truth (GT) aspect and its corresponding seed words are provided.

Figure 4: Micro-averaged F1 scores of (1) and (2) with different $d_1$, $d_2$, and $d_3$ on a) B/T and b) Boots datasets.

Then, we investigated the sensitivity of latent semantic modeling distance $d_1$, $d_2$, and $d_3$ on (1) and (2), as shown in Figure 4. We observed the following observations. First, both (1) and (2) achieve the best results when a small $d_1$, e.g., $d_1 \leq 8$, is set, demonstrating the importance of narrowing the gap between seed word pairs when modeling latent semantic meanings. Also, (1) and (2) both perform better when a large $d_2$, e.g., $d_2 \geq 64$, is set, verifying the importance of independence modeling. Last, (1) and (2) achieve the best performance when $d_3$ is set to around 8 to 32, perhaps due to the strong regularization on each latent semantic meaning introduced when $d_3$ is too large.

Case study

To more closely investigate the aspect inference ability of HDAE, we compare the predictions made by HDAE, MATE, and TS-W2V, the results of which are shown in Table 4. For the example in Table 4(b), we see that the review segment contains keywords such as color and blue which are explicitly captured in aspect seed words. All models correctly infer and review the segment’s aspect. However, for cases in Table 4(c,d,e), the reviews’ segments do not explicitly match their aspect seed words but instead match the hyponymic relations (is-a) present between seed words and review segments. For example, there are hierarchical relations such as grayish brown is a color, leather is a material, and stiff is a type of difficult for cases in Table 4(c,d,e). We find only HDAE correctly recognizes the review segments’ aspects. We thus conclude HDAE captures and utilizes hyponymic relations (is-a) present between seed words and review segments, deriving reasonable aspect inference for each review segment and thus achieving better performance. Analogous behavior is observed for other cases in the appendix.

To explore the interpretability of the seed words’ latent semantic meanings, we conducted a case study in which we randomly selected review segments from the boot domain’s look aspect and investigated its association with each aspect of latent semantic meaning. Figure 5 shows the review segments captured by each seed word’s latent semantic meaning: we find that each aspect’s latent semantics focus on a distinct type of review segment. For example, for the seed word design, the latent semantic meaning $s_{d_1}^{j}$ focuses on segments with color information, whereas $s_{d_2}^{j}$ focuses on segments with the great keyword. Likewise, for the seed word attractive, the latent semantic meaning $s_{i,j}^{d_1}$ focuses on segments with cute information, whereas $s_{i,j}^{d_3}$ focuses on segments with unattractive information. These results demonstrate that the proposed aspect disentanglement module assists HDAE in modeling different latent semantics for each seed word. Also, HDAE finds the most relevant latent semantic meanings for each review segment, explaining the improvements in the aspect inference ability.

Conclusions and Future Work

We present HDAE, a hyperbolic disentangled aspect extractor which includes a hyperbolic aspect classifier and an aspect disentanglement module. On two datasets, HDAE, with its 97.8% reductions in parameter size versus TS-BT, shows superior aspect inference ability, further substantiated by an embedding visualization. The effect of the proposed components is proven by an ablation study, a parameter sensitivity study, and a case study.

In the future, we plan to explore the proposed module on...
other aspect-based sentiment analysis (ABSA) subtasks. We would also like to further improve the performance of the proposed components, for instance by setting up alignment scores for different aspect word pairs when modeling seed word dependence.

Acknowledgement

This research is partially supported by Ministry of Science and Technology, Taiwan under the project contract 110-2221-E-001-001- and 110-2634-F-002-051.

References

Aly, R.; Acharya, S.; Ossa, A.; Köhn, A.; Biemann, C.; and Panchenko, A. 2019. Every child should have parents: A taxonomy refinement algorithm based on hyperbolic term embeddings. arXiv preprint arXiv:1906.02002.

Angelidis, S.; and Lapata, M. 2018. Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. arXiv preprint arXiv:1808.08858.

Balazevic, I.; Allen, C.; and Hospedales, T. M. 2019. Multi-relational Poincaré Graph Embeddings. CoRR, abs/1905.09791.

Chami, I.; Wolf, A.; Juan, D.; Sala, F.; Ravi, S.; and Ré, C. 2020. Low-Dimensional Hyperbolic Knowledge Graph Embeddings. CoRR, abs/2005.00545.

Chami, I.; Ying, R.; Ré, C.; and Leskovec, J. 2019. Hyperbolic Graph Convolutional Neural Networks. CoRR, abs/1910.12933.

Chen, M.; Tang, Q.; Livescu, K.; and Gimpel, K. 2019. Variational sequential labelers for semi-supervised learning. arXiv preprint arXiv:1906.09535.

Chen, Z.; Mukherjee, A.; and Liu, B. 2014. Aspect extraction with automated prior knowledge learning. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 347–358.

Chin, J. Y.; Zhao, K.; Joty, S.; and Cong, G. 2018. ANR: Aspect-based neural recommender. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management, 147–156.

Dupont, E. 2018. Learning Disentangled Joint Continuous and Discrete Representations. arXiv:1804.00104.

Feng, V. W.; and Hirst, G. 2012. Text-level Discourse Parsing with Rich Linguistic Features. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 60–68. Jeju Island, Korea: Association for Computational Linguistics.

García-Pablos, A.; Cuadros, M.; and Rigau, G. 2018. W2VLD: Almost unsupervised system for aspect based sentiment analysis. Expert Systems with Applications, 91: 127–137.

Gülçehre, Ç.; Denil, M.; Malinowski, M.; Razavi, A.; Pascanu, R.; Hermann, K. M.; Battaglia, P. W.; Bapst, V.; Raposo, D.; Santoro, A.; and de Freitas, N. 2018. Hyperbolic Attention Networks. CoRR, abs/1805.09786.

He, R.; Lee, W. S.; Ng, H. T.; and Dahlmeier, D. 2017. An unsupervised neural attention model for aspect extraction. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 388–397.

He, X.; and Chua, T.-S. 2017. Neural factorization machines for sparse predictive analytics. In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, 355–364.

Hou, S.; Fan, Y.; Ju, M.; Ye, Y.; Wan, W.; Wang, K.; Mei, Y.; Xiong, Q.; and Shao, F. 2021. Disentangled Representation Learning in Heterogeneous Information Network for Large-scale Android Malware Detection in the COVID-19 Era and Beyond. 35th AAAI Conference on Artificial Intelligence (AAAI).

Hu, L.; Xu, S.; Li, C.; Yang, C.; Shi, C.; Duan, N.; Xie, X.; and Zhou, M. 2020. Graph Neural News Recommendation with Unsupervised Preference Disentanglement. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 4255–4264. Online: Association for Computational Linguistics.

Hu, M.; and Liu, B. 2004. Mining and summarizing customer reviews. In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 168–177.

Hu, Z.; Yang, Z.; Liang, X.; Salakhutdinov, R.; and Xing, E. P. 2017. Controllable Text Generation. CoRR, abs/1703.00955.

Huang, J.; Wang, Z.; Chang, K. C.; Hwu, W.; and Xiong, J. 2020. Exploring Semantic Capacity of Terms. CoRR, abs/2010.01898.

Iyyer, M.; Guha, A.; Chaturvedi, S.; Boyd-Graber, J.; and Daumé III, H. 2016. Feeding families and former friends: Unsupervised learning for dynamic fictional relationships. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1534–1544.

Jang, E.; Gu, S.; and Poole, B. 2017. Categorical Reparameterization with Gumbel-Softmax.

Jin, W.; Ho, H. H.; and Srirhari, R. K. 2009. OpinionMiner: A novel machine learning system for web opinion mining and extraction. In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1195–1204.

Karmanolakis, G.; Hsu, D.; Hsu, D.; and Gravano, L. 2019. Leveraging Just a Few Keywords for Fine-Grained Aspect Detection Through Weakly Supervised Co-Training. CoRR, abs/1909.00415.

Kingma, D. P.; and Ba, J. 2014. Adam: A Method for Stochastic Optimization. arXiv e-prints, arXiv:1412.6980.

Krioukov, D.; Papadopoulos, F.; Kitsak, M.; Vahdat, A.; and Boguñá, M. 2010. Hyperbolic geometry of complex networks. Physical Review E, 82(3).

Le, M.; Roller, S.; Papaxanthos, L.; Kiela, D.; and Nickel, M. 2019. Inferring Concept Hierarchies from Text Corpora via Hyperbolic Embeddings. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 3231–3241.
Liu, B. 2012. Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, 5(1): 1–167.

Liu, P.; Joty, S.; and Meng, H. 2015. Fine-grained opinion mining with recurrent neural networks and word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1433–1443.

Liu, Q.; Nickel, M.; and Kiela, D. 2019. Hyperbolic Graph Neural Networks. *CoRR*, abs/1910.12892.

Liu, Y.; Wang, X.; Wu, S.; and Xiao, Z. 2020. Independence Promoted Graph Disentangled Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04): 4916–4923.

López, F.; Heinzerling, B.; and Strube, M. 2019. Fine-Grained Entity Typing in Hyperbolic Space. *CoRR*, abs/1906.02505.

López, F.; and Strube, M. 2020. A Fully Hyperbolic Neural Model for Hierarchical Multi-Class Classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 460–475. Online: Association for Computational Linguistics.

Lund, J.; Cook, C.; Seppi, K.; and Boyd-Graber, J. 2017. Tandem anchoring: A multiword anchor approach for interactive topic modeling. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 896–905.

Luo, L.; Ao, X.; Song, Y.; Li, J.; Yang, X.; He, Q.; and Yu, D. 2019. Unsupervised Neural Aspect Extraction with Sememes. In *IJCAI*, 5123–5129.

Ma, J.; Cui, P.; Kuang, K.; Wang, X.; and Zhu, W. 2019a. Disentangled graph convolutional networks. In *International Conference on Machine Learning*, 4212–4221. PMLR.

Ma, J.; Zhou, C.; Cui, P.; Yang, H.; and Zhu, W. 2019b. Learning Disentangled Representations for Recommendation. *CoRR*, abs/1910.14238.

Mitchell, M.; Aguilar, J.; Wilson, T.; and Van Durme, B. 2013. Open domain targeted sentiment. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1643–1654.

Murty, S.; Verga, P.; Vilnis, L.; Radovanovic, I.; and McCallum, A. 2018. Hierarchical losses and new resources for fine-grained entity typing and linking. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 97–109.

Ni, J.; Li, J.; and McAuley, J. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, 188–197.

Nickel, M.; and Kiela, D. 2017. Poincaré embeddings for learning hierarchical representations. *arXiv preprint arXiv:1705.08039*.

Nickel, M.; and Kiela, D. 2018. Learning Continuous Hierarchies in the Lorentz Model of Hyperbolic Geometry. *CoRR*, abs/1806.03417.

Pontiki, M.; Galanis, D.; Papageorgiou, H.; Androutsopoulos, I.; Manandhar, S.; Al-Smadi, M.; Al-Ayyoub, M.; Zhao, Y.; Qin, B.; De Clercq, O.; et al. 2016. SemEval-2016 Task 5: Aspect based sentiment analysis. In *International Workshop on Semantic Evaluation*, 19–30.

Robbin, J. W.; and Salamon, D. A. 2011. Introduction to differential geometry. *ETH, Lecture Notes, preliminary version*, 18.

Ruder, S.; Ghaffari, P.; and Breslin, J. G. 2016. A Hierarchical Model of Reviews for Aspect-based Sentiment Analysis. *CoRR*, abs/1609.02745.

Sa, C. D.; Gu, A.; Ré, C.; and Sala, F. 2018. Representation Tradeoffs for Hyperbolic Embeddings. *CoRR*, abs/1804.03329.

Sala, F.; De Sa, C.; Gu, A.; and Ré, C. 2018. Representation tradeoffs for hyperbolic embeddings. In *International Conference on Machine Learning*, 4460–4469. PMLR.

Sarkar, R. 2011. Low distortion Delaunay embedding of trees in hyperbolic plane. In *International Symposium on Graph Drawing*, 355–366. Springer.

Shen, T.; Lei, T.; Barzilay, R.; and Jaakkola, T. 2017. Style transfer from non-parallel text by cross-alignment. *arXiv preprint arXiv:1705.09655*.

Shi, T.; Kang, K.; Choo, J.; and Reddy, C. K. 2018. Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. In *Proceedings of the 2018 World Wide Web Conference*, 1105–1114.

Shi, T.; Li, L.; Wang, P.; and Reddy, C. K. 2020. A Simple and Effective Self-Supervised Contrastive Learning Framework for Aspect Detection. arXiv:2009.09107.

Srivastava, A.; and Sutton, C. 2017. Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*.

Sun, J.; Cheng, Z.; Zuberi, S.; Pérez, F.; and Volkovs, M. 2021. HGCF: Hyperbolic Graph Convolution Networks for Collaborative Filtering.

Tai, C.-Y.; Huang, C.-K.; Huang, L.-Y.; and Ku, L.-W. 2021. Knowledge Based Hyperbolic Propagation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1945–1949.

Tay, Y.; Luu, A. T.; and Hui, S. C. 2017. Enabling Efficient Question Answer Retrieval via Hyperbolic Neural Networks. *CoRR*, abs/1707.07847.

Tulkens, S.; and van Cranenburgh, A. 2020. Embarrassingly Simple Unsupervised Aspect Extraction. *CoRR*, abs/2004.13580.

Wang, W.; Pan, S. J.; Dahlmeier, D.; and Xiao, X. 2016. Recursive neural conditional random fields for aspect-based sentiment analysis. *arXiv preprint arXiv:1603.06679*.

Wang, X.; Jin, H.; Zhang, A.; He, X.; Xu, T.; and Chua, T.-S. 2020. Disentangled Graph Collaborative Filtering. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1001–1010.

Xu, P.; and Barbosa, D. 2018. Neural fine-grained entity type classification with hierarchy-aware loss. *arXiv preprint arXiv:1803.03378*. 
Yang, B.; and Cardie, C. 2012. Extracting opinion expressions with semi-Markov conditional random fields. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing, 1335–1345.

Yoshua Bengio, A. C. C.; and Vincent, P. 2012. Unsupervised Feature Learning and Deep Learning: A Review and New Perspectives. CoRR, abs/1206.5538.

Zhang, C.; and Gao, J. 2020. Hype-HAN: Hyperbolic Hierarchical Attention Network for Semantic Embedding. In IJCAI, 3990–3996.

Zhao, J.; Kim, Y.; Zhang, K.; Rush, A.; and LeCun, Y. 2018. Adversarially regularized autoencoders. In International Conference on Machine Learning.

Zhuang, H.; Guo, F.; Zhang, C.; Liu, L.; and Han, J. 2020. Joint Aspect-Sentiment Analysis with Minimal User Guidance. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 1241–1250.

**APPENDIX**

We provide more information on datasets (Section 1.1) and implementation details (Section 1.2). Besides, we report more detailed experimental results (Section 1.3), and parameter sensitivity analysis (Section 1.4).

**Datasets**

During training, segment aspect labels (9-class for product reviews and 12-class for restaurant reviews) are not available but provided during validation and test. For instance, in the Laptop Bags domain, the review segments’ aspects could be Compartments, Customer Service, Looks, or Price. Note that a general aspect is assigned if the segment doesn’t discuss any specific aspects. For each domain, we train our model on a training set only with seed words G via the teacher. For the aspect seed words, we follow [Angelidis and Lapata 2018; Karamanolakis et al. 2019] to use the same 30 seed words for two datasets. Besides, for two datasets, we follow Angelidis and Lapata 2018; Karamanolakis et al. 2019 to do data preprocessing, such as removing stop-words.

For the Amazon review dataset, the reviews of each domain are already segmented by [Angelidis and Lapata 2018], where they use a Rhetorical Structure Theory parser (Feng and Hirst 2012) to segment reviews into elementary discourse units (EDUs). Across domains, the average numbers of training, validation, and test segments are around 1 million, 700 segments, respectively. For restaurant reviews, the reviews of each language are already segmented into sentences. Across languages, the average number of training and test segments is around 2500 and 800 segments, respectively.

**Implementation details**

For HDAE, the details hyper-parameter settings are given in Table 5 and 6, which are determined by optimizing on a validation set. We also provide the parameter sensitivity experiment of latent semantic modeling distance $d_1$, $d_2$, and $d_3$, Grumbel-Softmax temperature $\tau$, ratio of distillation objective $\lambda$, number of latent semantic I in section 11. The total number of negative examples $k_n$ was set to 10. We followed the procedure in Angelidis and Lapata 2018 to set the 200-dimensional word embeddings for the Amazon product reviews and the 300-dimensional multilingual word2vec embeddings from Ruder, Ghaffari, and Breslin 2016 for restaurant reviews. For all models, the same 30 seed words were set per aspect. For HDAE, model parameters are optimized by using the Adam optimizer (Kingma and Ba 2014). For setting distillation objective, the teacher, a bag-of-word classifier, is implement, and we use iterative co-training to update each seed word’s predictive quality. For TS-*, we report the result from iterative co-training, and in each round, we divide the learning rate by 10. For SSCL-BT*, we use code provided in and conduct aspect mapping after training the teacher model. The smooth factor $\lambda$ is set to 0.5 and temperature is set to 1. For all models, the learning rate was selected from $[2 \times 10^{-4}, 1 \times 10^{-6}, 5 \times 10^{-7}, 5 \times 10^{-8}, 1 \times 10^{-8}]$. Other hyperparameters were optimized according to validation results. For each model, we repeat each experiment 5 times and report the average test performance with the parameter configuration that achieves the best validation performance.

**Experimental Results**

we provide the performance the unsupervised based method ABAE and SSCL on English Restaurant review, as shown in Table 1.

Then, we provide more results for seed word based approaches’ inference performance per aspect and their corresponding embedding visualization on Bags, Bluetooth Headsets, Boots, Keyboards, Televisions, and Vacuums (VCs) datasets, shown in Figure 6, 7, 8, 9, 10, 11, respectively.

---

Table 5: Hyper-parameter settings for the product review.

| Bags | KBs | Boots | B/T | TVs | VC |
|------|-----|-------|-----|-----|----|
| $\beta$ | $\lambda$ | $\tau$ | $d_1$ | $d_2$ | $d_3$ |
| 0.01 | 1 | 1e-3 | 1 | 128 | 8 |
| 0.02 | 2 | 1e-2 | 128 | 128 | 64 |

Table 6: Hyper-parameter settings for the restaurant review datasets.

| Restaurant | Domain |
|-----------|-------|
| $\beta$ | $\lambda$ | $\tau$ | $d_1$ | $d_2$ | $d_3$ |
| En | Sp | Fr | Ru | Du | Tur |
| 0.01 | 100 | 1000 | 10000 | 10000 | 10000 |
| 0.02 | 100 | 1000 | 10000 | 10000 | 10000 |
| 0.03 | 100 | 1000 | 10000 | 10000 | 10000 |
| 0.04 | 100 | 1000 | 10000 | 10000 | 10000 |
| 0.05 | 100 | 1000 | 10000 | 10000 | 10000 |

---

10https://github.com/tshi04/AspDecSSCL
Table 7: Micro-averaged F1 reported for 12-class sentence-level aspect detection in restaurant reviews.

| Model     | En | Sp | Fr | Ru | Du | Tur |
|-----------|----|----|----|----|----|-----|
| ABAE      | 35.8 | -  | -  | -  | -  | -   |
| SSCL-BT*  | 47.3 | -  | -  | -  | -  | -   |
| LDA-AR    | 28.5 | 17.7 | 13.1 | 14.8 | 25.9 | 27.7 |
| MATE      | 41.0 | 24.9 | 17.8 | 18.4 | 36.1 | 39.0 |
| MATE-UW   | 40.3 | 18.3 | 19.2 | 21.8 | 31.5 | 25.2 |
| TS-Teacher| 44.9 | 41.8 | 34.1 | 54.4 | 40.7 | 30.2 |
| TS-ATT    | 47.8 | 41.7 | 52.4 | 59.0 | 42.1 | 42.3 |
| TS-BT     | 51.8 | 42.0 | 39.2 | 58.0 | 43.0 | 45.0 |
| HDAE      | 57.9 | 65.7 | 48.6 | 62.9 | 57.2 | 50.8 |

Parameter Sensitivity Analysis

In this section, we provide more results for parameter sensitivity. Table 8 shows effects of the grumbel-softmax temperature $\tau$ on the performance of HDAE. We find that our model achieves the best results when small $\tau$ is set, suggesting that it is important to not to mix the latent semantic when predicting the segment’s aspect. Table 10 gives results for baselines with (HDAE, MATE, TS-ATT, TS-BT) or without (W2V, BERT) leveraging seed word in different ground truth aspect labels ratios.

Table 8: Micro-averaged F1 of HDAE given grumbel-softmax temperature $\tau$.

| $\tau$ | 1e-1 | 1e-2 | 1e-3 | 1e-4 | 1e-5 | 1e-6 |
|--------|------|------|------|------|------|------|
| Bags   | 67.1 | 67.5 | 68.8 | 67.8 | 67.8 | 68.2 |
| B/T    | 68.4 | 69.1 | 71.1 | 71.9 | 70.4 | 70.3 |
| Boots  | 61.8 | 62.4 | 63.5 | 63.7 | 64.0 | 63.7 |
| TVs    | 69.8 | 70.8 | 71.2 | 70.4 | 70.3 | 70.0 |

Table 9: Micro-averaged F1 of HDAE given $\lambda$.

| $\lambda$ | 0   | 5   | 10  | 100 | 500 | 1000 | 3000 |
|-----------|-----|-----|-----|-----|-----|------|------|
| Bags      | 67.3 | 68.8 | 68.1 | 67.5 | 67.0 | 66.9 | 66.0 |
| B/T       | 70.1 | 71.1 | 70.8 | 69.0 | 67.9 | 67.2 | 63.3 |
| Boots     | 60.5 | 62.4 | 63.7 | 62.8 | 62.4 | 61.2 | 60.7 |
| TVs       | 61.1 | 70.1 | 71.2 | 70.3 | 70.1 | 69.8 | 69.4 |

Table 10: Micro-averaged F1 of HDAE given $I$.

Results in ground truth aspect labels ratios

In this section, we provide results for baselines with (HDAE, MATE, TS-ATT, TS-BT) or without (W2V, BERT) leveraging seed word in different ground truth aspect labels ratios. We find the proposed model HDAE can achieve best performance in different ground truth aspect labels ratios, suggesting the effectiveness of purpose hyperbolic disentangled based method. Besides, we notice that in the low aspect labels data ratios, seed word based approaches (MATE, TS-ATT, TS-BT, and HDAE) can achieve better, demonstrating that seed words can give useful guidance and assist models to improve aspect inference ability.
Figure 6: Inference performance per aspect of HDAE, TS-W2V, and MATE on Bags dataset. The following figure is segment vector t-SNE visualization of each model, where the different color of point represent different aspect.

| β      | 0.005 | 0.01  | 0.02  | 0.05  |
|--------|-------|-------|-------|-------|
| Bags   | 67.8  | 68.0  | 67.7  | 66.6  |
| B/T    | 70.5  | 71.2  | 71.9  | 71.3  |
| Boots  | 63.1  | 64.0  | 63.7  | 63.2  |
| TVs    | 68.1  | 69.8  | 70.5  | 70.0  |

Table 11: Micro-averaged F1 of HDAE given β

| ratio r | Restaurant review domain (En) |
|---------|-------------------------------|
|         | 10%  | 30%  | 50%  | 70%  | 100% |
| W2V-Gold| 16.3 | 33.0 | 38.6 | 46.7 | 58.8 |
| BERT-Gold| 24.8 | 36.5 | 48.5 | 55.9 | 63.1 |
| MATE    | 43.8 | 46.7 | 50.4 | 54.3 | 60.1 |
| TS-ATT  | 48.5 | 50.6 | 53.2 | 57.7 | 61.1 |
| TS-BT   | 53.6 | 56.1 | 58.4 | 61.2 | 64.2 |
| HDAE    | 58.6 | 62.2 | 64.1 | 66.9 | 70.5 |

Table 12: Micro-averaged F1 for 12-class sentence-level aspect detection in restaurant reviews in English with different ratios of training set r.

| ratio r | Restaurant review domain (Fr) |
|---------|-------------------------------|
|         | 10%  | 30%  | 50%  | 70%  | 100% |
| W2V-Gold| 21.3 | 29.9 | 37.1 | 43.1 | 50.4 |
| BERT-Gold| 20.2 | 24.8 | 33.0 | 40.9 | 50.6 |
| MATE    | 28.7 | 34.8 | 40.5 | 45.2 | 48.1 |
| TS-ATT  | 32.8 | 38.1 | 44.1 | 46.6 | 50.1 |
| TS-BT   | 43.0 | 44.9 | 46.5 | 48.0 | 53.0 |
| HDAE    | 48.7 | 51.8 | 54.9 | 60.8 | 65.4 |

Table 13: Micro-averaged F1 for 12-class sentence-level aspect detection in restaurant reviews in French with different ratios of training set r.

| ratio r | Restaurant review domain (Ru) |
|---------|-------------------------------|
|         | 10%  | 30%  | 50%  | 70%  | 100% |
| W2V-Gold| 21.2 | 29.9 | 37.1 | 43.1 | 50.4 |
| BERT-Gold| 23.5 | 31.5 | 41.3 | 47.9 | 55.3 |
| MATE    | 22.6 | 30.3 | 36.8 | 44.3 | 51.2 |
| TS-ATT  | 58.8 | 59.1 | 59.8 | 62.1 | 65.5 |
| TS-BT   | 59.5 | 61.3 | 62.1 | 65.5 | 67.4 |
| HDAE    | 61.3 | 65.0 | 67.8 | 71.5 | 76.8 |

Table 14: Micro-averaged F1 for 12-class sentence-level aspect detection in restaurant reviews in Russian with different ratios of training set r.
Figure 7: Inference performance per aspect of HDAE, TS-W2V, and MATE on Keyboards dataset. The following figure is segment vector t-SNE visualization of each model, where the different color of point represent different aspect.

Table 16: Micro-averaged F1 for 12-class sentence-level aspect detection in restaurant reviews in Dutch with different ratios of training set r.

| ratio r | Restaurant review domain (Du) | 10% | 30% | 50% | 70% | 100% |
|---------|------------------------------|-----|-----|-----|-----|------|
| W2V-Gold |                              | 24.4| 32.8| 42.4| 47.3| 51.4 |
| BERT-Gold |                            | 28.1| 40.8| 47.5| 51.6| 53.5 |
| MATE     |                              | 38.8| 45.8| 51.8| 53.5| 55.4 |
| TS-ATT   |                              | 43.9| 47.1| 50.4| 53.6| 57.6 |
| TS-BT    |                              | 45.4| 47.1| 50.4| 53.6| 57.6 |
| HDAE     |                              | 58.5| 62.5| 68.3| 72.1| 73.8 |

Table 17: Micro-averaged F1 for 12-class sentence-level aspect detection in restaurant reviews in Turkish with different ratios of training set r.

| ratio r | Restaurant review domain (Tur) | 10% | 30% | 50% | 70% | 100% |
|---------|------------------------------|-----|-----|-----|-----|------|
| W2V-Gold |                              | 28.6| 37.2| 42.3| 50.8| 55.7 |
| BERT-Gold |                            | 31.5| 39.0| 45.6| 52.3| 56.5 |
| MATE     |                              | 41.3| 44.9| 47.1| 49.9| 53.0 |
| TS-ATT   |                              | 45.9| 47.5| 49.7| 52.8| 55.5 |
| TS-BT    |                              | 45.5| 48.7| 52.6| 54.3| 57.6 |
| HDAE     |                              | 49.8| 52.4| 56.9| 60.1| 65.4 |

Table 18: Micro-averaged F1 of HDAE given # of seed words

| # of seed words | 0  | 5  | 15 | 20 | 30  |
|-----------------|----|----|----|----|-----|
| Bags            | 41.2| 61.7| 63.6| 67.2| 68.8 |
| KBs             | 33.2| 65.2| 68.2| 69.4| 72.2 |
| B/T             | 42.3| 70.2| 70.5| 71.9| 72.0 |
| Boots           | 37.2| 61.0| 63.5| 63.2| 64.0 |
| TV              | 45.1| 66.3| 65.7| 68.2| 71.2 |
| VCs             | 40.2| 59.6| 61.2| 66.1| 66.9 |

Table 19: Micro-averaged F1 of HDAE given ratio of $J_{d_1}$

| ratio of $J_{d_1}$ | 0.5 | 1   | 5   | 10  | 100 |
|--------------------|-----|-----|-----|-----|-----|
| Bags               | 67.9| 68.8| 68.1| 66.9| 66.1 |
| KBs                | 71.9| 72.2| 71.3| 71.0| 70.6 |
| B/T                | 72.2| 72.0| 71.7| 71.2| 71.0 |
| Boots              | 64.8| 64.0| 64.1| 63.3| 61.5 |
| TV                 | 69.2| 71.2| 70.3| 69.5| 68.9 |
| VCs                | 65.0| 66.9| 66.7| 64.2| 63.9 |

Table 20: Micro-averaged F1 of HDAE given ratio of $J_{d_2}$

| ratio of $J_{d_2}$ | 0.5 | 1   | 5   | 10  | 100 |
|--------------------|-----|-----|-----|-----|-----|
| Bags               | 67.9| 68.8| 68.1| 67.8| 67.2 |
| KBs                | 71.0| 72.2| 71.9| 71.0| 71.3 |
| B/T                | 72.2| 72.0| 72.0| 71.3| 70.9 |
| Boots              | 64.8| 64.0| 63.5| 62.3| 62.5 |
| TV                 | 69.2| 71.2| 69.2| 68.5| 68.7 |
| VCs                | 65.0| 66.9| 65.8| 65.3| 64.1 |

Table 21: Micro-averaged F1 of HDAE given ratio of $J_{d_3}$
Figure 8: Inference performance per aspect of HDAE, TS-W2V, and MATE on Boots dataset. The following figure is segment vector t-SNE visualization of each model, where the different color of point represent different aspect.

Figure 9: Inference performance per aspect of HDAE, TS-W2V, and MATE on Bluetooth Headsets dataset. The following figure is segment vector t-SNE visualization of each model, where the different color of point represent different aspect.
Figure 10: Inference performance per aspect of HDAE, TS-W2V, and MATE on Televisions dataset. The following figure is segment vector t-SNE visualization of each model, where the different color of point represent different aspect.

Figure 11: Inference performance per aspect of HDAE, TS-W2V, and MATE on Vacuums dataset. The following figure is segment vector t-SNE visualization of each model, where the different color of point represent different aspect.
| Do not purchase. GT: Noise |
|---------------------------|
| Seed Words: loud, noise, noisy, quiet, action, sound, quieter, know, make |
| HDAE: General ✗ MATE: General ✗ TS-W2V: General ✗ |

| Which died. GT: General |
|-------------------------|
| Seed Words: think, recommend, purchase, using, unit, star, microsoft, mouse |
| HDAE: Build Quality ✗ MATE: Build Quality ✗ TS-W2V: Build Quality ✗ |

| Except the keyboard was one of those high keyed, clackety-clunkety types. GT: General |
|------------------|
| Seed Words: think, recommend, purchase, using, unit, star, microsoft, mouse |
| HDAE: General ✓ MATE: General ✓ TS-W2V: Comfort ✓ |

| I really liked the look of it. GT: Looks |
|------------------|
| Seed Words: look, slim, original, appearance, little, attractive, beautiful |
| HDAE: Looks ✓ MATE: General ✓ TS-W2V: Looks ✓ |

| I liked the feel of the keys. GT: Comfort |
|------------------|
| Seed Words: feel, comfortable, mushy, key, like, keyboard, good, perfect |
| HDAE: Comfort ✓ MATE: General ✓ TS-W2V: Looks ✓ |

| But it has all the buttons to interface with my iMac. GT: Extra functionality |
|------------------|
| Seed Words: buttons, light, pencil, volume, power, feature, bright, mute, handy, low, dark |
| HDAE: Extra functionality ✓ MATE: Extra functionality ✓ TS-W2V: Extra functionality ✓ |

| It is quiet. GT: Noise |
|------------------|
| Seed Words: loud, noise, noisy, red, action, sound, quieter, know, make |
| HDAE: Noise ✓ MATE: Noise ✓ TS-W2V: Noise ✓ |

| The layout of the keys makes it difficult for me to use, with keys like the backspace. GT: Layout |
|------------------|
| Seed Words: key, delete, backspace, size, layout, end, insert, home, bar, perfect, space |
| HDAE: Layout ✓ MATE: General ✓ TS-W2V: Layout ✓ |

| And doesn’t depress at times GT: Build Quality |
|------------------|
| Seed Words: working, build, stopped, quality, month, spacebar, stuck, left, plastic, kind, died |
| HDAE: Build Quality ✓ MATE: Build Quality ✓ TS-W2V: General ✓ |

| And the key for the “ t ” is already broken. GT: Build Quality |
|------------------|
| Seed Words: working, build, stopped, quality, month, spacebar, stuck, left, plastic, kind, died |
| HDAE: Build Quality ✓ MATE: Layout ✗ TS-W2V: Comfort ✓ |

| Has a top row of quick link GT: Extra functionality |
|------------------|
| Seed Words: button, light, pencil, volume, power, feature, bright, mute, handy, low, dark |
| HDAE: Extra functionality ✓ MATE: Extra functionality ✓ TS-W2V: Comfort ✓ |

| The keyboard is sleek and visually appealing. GT: Looks |
|------------------|
| Seed Words: look, slim, original, appearance, little, attractive, beautiful |
| HDAE: Looks ✓ MATE: General ✓ TS-W2V: General ✓ |

| That had popped off. GT: Build Quality |
|------------------|
| Seed Words: working, build, stopped, quality, month, spacebar, stuck, left, plastic, kind, died |
| HDAE: Build Quality ✓ MATE: Build Quality ✓ TS-W2V: General ✓ |

| It is very responsive. GT: Comfort |
|------------------|
| Seed Words: feel, comfortable, mushy, key, like, keyboard, good, perfect, press, wrist, action, shallow, smooth |
| HDAE: Comfort ✓ MATE: Connectivity ✗ TS-W2V: General ✓ |

Table 22: Comparison of predictions on sample Keyboards product review segments between HDAE, MATE, and TS-W2V. For each review segment, the ground true (GT) aspect and its corresponding seed words are provided.
Table 23: Comparison of predictions on sample Boots product review segments between HDAE, MATE, and TS-W2V. For each review segment, the ground true (GT) aspect and its corresponding seed words are provided.
Table 24: Comparison of predictions on sample Bags product review segments between HDAE, MATE, and TS-W2V. For each review segment, the ground true (GT) aspect and its corresponding seed words are provided.
Table 25: Comparison of predictions on sample Televisions product review segments between HDAE, MATE, and TS-W2V. For each review segment, the ground true (GT) aspect and its corresponding seed words are provided.

| GT: Apps Interface | HDAE: Ease of use | MATE: Size Look | TS-W2V: General |
|--------------------|-------------------|-----------------|-----------------|
| The Yahoo! widgets do not work. | Seed Words: netflix, user, file, hulu, apps, watch, flash, internet, smart, video |
| HDAE: Apps Interface  | MATE: Customer Service  | TS-W2V: Apps Interface  |
| The picture quality is very sharp and crisp | Seed Words: picture, color, quality, back, bright, nice, clear, look, excellent, crisp, screen, right, dead, pixel, trace, beautiful |
| The price is enticing. | GT: Price |
| But bright colors generally looked | Seed Words: picture, color, quality, back, bright, nice, clear, look, excellent, crisp, screen, right, dead, pixel, trace, beautiful |
| The sound from the TV itself is very tiny. | GT: Sound |
| The cable connection port | GT: Connectivity |
| Picture and sound are both acceptable. | GT: Sound |
| Are difficult to use and setup | GT: Ease of Use |
| Fast moving objects were incredibly pixelated and blotchy. | GT: Image |
| Lacks a net browser and the only thing | GT: Apps Interface |
| Is pandora. | GT: Apps Interface |
| Washed out but 720p movies. | GT: Image |
| HDAE: Image  | MATE: General  | TS-W2V: Connectivity  |

Seed Words: netflix, user, file, hulu, apps, watch, flash, internet, smart, video
Table 26: Comparison of predictions on sample Bluetooth Headsets product review segments between HDAE, MATE, and TS-W2V. For each review segment, the ground true (GT) aspect and its corresponding seed words are provided.
Eventually I started listening to my iPod.
Seed Words: vac, cleaner, vacuum, buy, bought, new, better, year, recommend, product, owned, review, gave, away, kenmore, dyson

| HDAE: Build Quality | MATE: Build Quality | TS-W2V: General |
|---------------------|---------------------|-----------------|
| It is easy to move because of the adjustable wheels on the side of the brush | GT: Ease of use |
| Seed Words: easy, cord, push, corner, vacuuming, pile, maneuver, nozzle, awkward, crevice, constantly, bog, impossible, short |
| HDAE: Ease of use | MATE: Accessories | TS-W2V: Ease of use |
| Too bulky, cord is unusually stiff and tangles are impossible to remove. | GT: Ease of use |
| Seed Words: easy, cord, push, corner, vacuuming, pile, maneuver, nozzle, awkward, crevice, constantly, bog, impossible, short |
| HDAE: Ease of use | MATE: Build Quality | TS-W2V: Ease of use |
| It is so easy to maneuver. | GT: Ease of Use |
| Seed Words: easy, cord, push, corner, vacuuming, pile, maneuver, nozzle, awkward, crevice, constantly, bog, impossible, short |
| HDAE: Ease of use | MATE: Weight | TS-W2V: Ease of Use |
| Because it was so loud. | GT: Noise |
| Seed Words: quiet, noisy, loud, powerful, noise, louder, ear, loudest, light, incredibly, deafening, seriously, actually |
| HDAE: Noise | MATE: Noise | TS-W2V: Noise |
| And have powerful suction. | GT: Suction Power |
| Seed Words: suction, pick, powerful, power, good, hair, carpet, such, quiet, really, performs, dirt, tile, ok |
| HDAE: Suction Power | MATE: Suction Power | TS-W2V: Suction Power |
| While the suction is very good. | GT: Suction Power |
| Seed Words: suction, pick, powerful, power, good, hair, carpet, such, quiet, really, performs, dirt, tile, ok |
| HDAE: Suction Power | MATE: Noise | TS-W2V: Suction Power |
| Which prevents wearing and tearing the plastic parts on the brush | GT: Build Quality |
| Seed Words: belt, broke, turn, working, burning, electrical, built, stop, month, roller, time, minute |
| HDAE: Build Quality | MATE: Accessories | TS-W2V: Build Quality |
| The sides wrapped with protective rubber like rim | GT: Build Quality |
| Seed Words: belt, broke, turn, working, burning, electrical, built, stop, month, roller, time, minute, problem |
| HDAE: Build Quality | MATE: Noise | TS-W2V: General |
| Then the engine completely stopped vacuuming. | GT: Build Quality |
| Seed Words: belt, broke, turn, working, burning, electrical, built, stop, month, roller, time, minute, problem, brush, design |
| HDAE: Build Quality | MATE: Suction Power | TS-W2V: Ease of Use |
| A small, light-weight appliance that can do a big job. | GT: Weight |
| Seed Words: light, weight, lightweight, heavy, size, compact, maneuver, guess, quiet, quite, probably |
| HDAE: Weight | MATE: Price | TS-W2V: Suction Power |
| You have to hold it at a very uncomfortable angle | GT: Ease of Use |
| Seed Words: easy, cord, push, corner, vacuuming, pile, maneuver, nozzle, awkward, crevice, constantly, bog, impossible, short |
| HDAE: Ease of Use | MATE: Ease of Use | TS-W2V: General |
| The tools were previously stored inside the canister | GT: Accessories |
| Seed Words: filter, brush, attachment, roll, turbo, easily, expensive, wide, turn, bag, replacing, typical, hepa |
| HDAE: Accessories | MATE: Build Quality | TS-W2V: General |

Table 27: Comparison of predictions on sample Vacuums product review segments between HDAE, MATE, and TS-W2V. For each review segment, the ground true (GT) aspect and its corresponding seed words are provided.