Leveraging Just a Few Keywords for Fine-Grained Aspect Detection Through Weakly Supervised Co-Training

Giannis Karamanolakis, Daniel Hsu, Luis Gravano
Columbia University, New York, NY 10027, USA
{gkaraman, djhsu, gravano}@cs.columbia.edu

Abstract

User-generated reviews can be decomposed into fine-grained segments (e.g., sentences, clauses), each evaluating a different aspect of the principal entity (e.g., price, quality, appearance). Automatically detecting these aspects can be useful for both users and downstream opinion mining applications. Current supervised approaches for learning aspect classifiers require many fine-grained aspect labels, which are labor-intensive to obtain. And, unfortunately, unsupervised topic models often fail to capture the aspects of interest. In this work, we consider weakly supervised approaches for training aspect classifiers that only require the user to provide a small set of seed words (i.e., weakly positive indicators) for the aspects of interest. First, we show that current weakly supervised approaches do not effectively leverage the predictive power of seed words for aspect detection. Next, we propose a student-teacher approach that effectively leverages seed words in a bag-of-words classifier (teacher); in turn, we use the teacher to train a second model (student) that is potentially more powerful (e.g., a neural network that uses pre-trained word embeddings). Finally, we show that iterative co-training can be used to cope with noisy seed words, leading to both improved teacher and student models. Our proposed approach consistently outperforms previous weakly supervised approaches (by 14.1 absolute F1 points on average) in six different domains of product reviews and six multilingual datasets of restaurant reviews.

1 Introduction

A typical review of an entity on platforms such as Yelp and Amazon discusses multiple aspects of the entity (e.g., price, quality) in individual review segments (e.g., sentences, clauses). Consider for example the Amazon product review in Figure 1. The text discusses various aspects of the TV such as price, ease of use, and sound quality. Given the vast number of online reviews, both sellers and customers would benefit from automatic methods for detecting fine-grained segments that discuss particular aspects of interest. Fine-grained aspect detection is also a key task in downstream applications such as aspect-based sentiment analysis and multi-document summarization (Hu and Liu, 2004; Liu, 2012; Pontiki et al., 2016; Angelidis and Lapata, 2018).

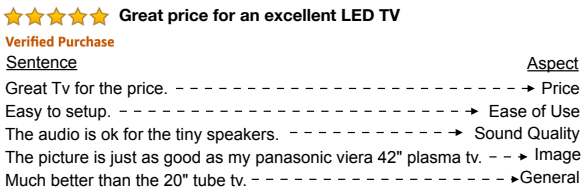

Figure 1: Example of product review with aspect annotations: each individual sentence of the review discusses a different aspect (e.g., price) of the TV.

In this work, we consider the problem of classifying individual segments of reviews to predefined aspect classes when ground truth aspect labels are not available. Indeed, reviews are often entered as unstructured, free-form text and do not come with aspect labels. Also, it is infeasible to manually obtain segment annotations for retail stores like Amazon with millions of different products. Unfortunately, fully supervised neural networks cannot be applied without aspect labels. Moreover, the topics learned by unsupervised neural topic models are not perfectly aligned with the users' aspects of interest, so substantial human effort is required for interpreting and mapping the learned topics to meaningful aspects.

Here, we investigate whether neural networks can be effectively trained under this challenging setting when only a small number of descriptive keywords, or *seed words*, are available for each

Aspect	Seed Words
Price (EN)	price, value, money, worth, paid
Image (EN)	picture, color, quality, black, bright
Food (EN)	food, delicious, pizza, cheese, sushi
Drinks (FR)	vin, bière, verre, bouteille, cocktail
Ambience (SP)	ambiente, mesas, terraza, acogedor, ruido

Table 1: Examples of aspects and five of their corresponding seed words in various domains (electronic products, restaurants) and languages ("EN" for English, "FR" for French, "SP" for Spanish).

aspect class. Table 1 shows examples of aspects and five of their corresponding seed words from our experimental datasets (described later in more detail). In contrast to a classification label, which is only relevant for a single segment, a seed word can implicitly provide aspect supervision to potentially many segments. We assume that the seed words have already been collected either manually or automatically. Indeed, collecting a small[1] set of seed words per aspect is typically easier than manually annotating thousands of segments for training neural networks. As we will see, even noisy seed words that are only weakly predictive of the aspect will be useful for aspect detection.

Training neural networks for segment-level aspect detection using just a few seed words is a challenging task. Indeed, as a contribution of this paper, we observe that current weakly supervised networks do not effectively leverage the predictive power of the available seed words. To address the shortcomings of previous seed word-based approaches, we propose a novel *weakly supervised* approach, which uses the available seed words in a more effective way. In particular, we consider a *student-teacher* framework, according to which a bag-of-seed-words classifier (teacher) is applied on unlabeled segments to supervise a second model (student), which can be any supervised model, including neural networks.

Our approach introduces several important contributions. First, our teacher model considers each individual seed word as a (noisy) aspect indicator, which as we will show, is more effective than previously proposed weakly supervised approaches. Second, by using only the teacher's aspect probabilities, our student generalizes better than the teacher and, as a result, the student outperforms both the teacher and previously proposed weakly supervised models. Finally, we show how iterative co-training can be used to cope with noisy seed words: the teacher effectively estimates the predictive quality of the noisy seed words in an unsupervised manner using the associated predictions by the student. Iterative co-training then leads to both improved teacher and student models. Overall, our approach consistently outperforms existing weakly supervised approaches, as we show with an experimental evaluation over six domains of product reviews and six multilingual datasets of restaurant reviews.

The rest of this paper is organized as follows. In Section 2 we review relevant work. In Section 3 we describe our proposed weakly supervised approach. In Section 4 we present our experimental setup and findings. Finally, in Section 5 we conclude and suggest future work. A preliminary version of this work was presented at the Second Learning from Limited Labeled Data Workshop (Karamanolakis et al., 2019).

2 Related Work and Problem Definition

We now review relevant work on aspect detection (Section 2.1), co-training (Section 2.2), and knowledge distillation (Section 2.3). We also define our problem of focus (Section 2.4).

2.1 Segment-Level Aspect Detection

The goal of segment-level aspect detection is to classify a segment s to K aspects of interest.

Supervised Approaches. Rule-based or traditional learning models for aspect detection have been outperformed by supervised neural networks (Liu et al., 2015; Poria et al., 2016; Zhang et al., 2018). Supervised neural networks first use an embedding function[2] (EMB) to compute a low dimensional segment representation $h = \mathrm{EMB}(s) \; 2 \; \mathrm{R}^d$ and then feed h to a classification layer[3] (CLF) to predict probabilities for the K aspect classes of interest: $p = hp^1; : : : ; p^K i = \mathrm{CLF}(h)$. For simplicity, we write $p = f(s)$. The parameters of the embedding function and the classification layer are learned using ground truth,

[1] In our experiments, we only consider around 30 seed words per aspect. For comparison, the vocabulary of the datasets has more than 10,000 terms.

[2] Examples of segment embedding functions are the average of word embeddings (Wieting et al., 2015; Arora et al., 2017), Recurrent Neural Networks (RNNs) (Yang et al., 2016; Wieting and Gimpel, 2017), Convolutional Neural Networks (CNNs) (Kim, 2014), self-attention blocks (Devlin et al., 2019; Radford et al., 2018), etc.

[3] The classification layer is usually a hidden layer followed by the softmax function.

segment-level aspect labels. However, aspect labels are not available in our setting, which hinders the application of supervised learning approaches.

Unsupervised Approaches. Topic models have been used to train aspect detection with unannotated documents. Recently, neural topic models (Iyyer et al., 2016; Srivastava and Sutton, 2071; He et al., 2017) have been shown to produce more coherent topics than earlier models such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003). In their Aspect Based Autoencoder (ABAE), He et al. (2017) first use segment s to predict aspect probabilities $p = f(s)$ and then use p to reconstruct an embedding h^0 for s as a convex combination of K aspect embeddings: $h^0 = \sum_{k=1}^{K} p^k A_k$, where $A_k \in \mathbb{R}^d$ is the embedding of the k-th aspect. The aspect embeddings A_k are initialized by clustering the vocabulary embeddings using k-means with K clusters. ABAE is trained by minimizing the segment reconstruction error.[4]

Unfortunately, unsupervised topic models are not effective when used directly for aspect detection. In particular, in ABAE, the K topics learned to reconstruct the segments are not necessarily aligned with the K aspects of interest. A possible fix is to first learn $K^0 >> K$ topics and do a K^0-to-K mapping as a post-hoc step. However, this mapping requires either aspect labels or substantial human effort for interpreting topics and associating them with aspects. This mapping is nevertheless not possible if the learned topics are not aligned with the aspects.

Weakly Supervised Approaches. Weakly supervised approaches use minimal domain knowledge (instead of ground truth labels) to model meaningful aspects. In our setting, domain knowledge is given as a set of seed words for each aspect of interest (Lu et al., 2011; Lund et al., 2017; Angelidis and Lapata, 2018). Lu et al. (2011) use seed words as asymmetric priors in probabilistic topic models (including LDA). Lund et al. (2017) use LDA with fixed topic-word distributions, which are learned using seed words as "anchors" for topic inference (Arora et al., 2013). Neither of these two approaches can be directly applied into more recent neural networks for aspect detection. Angelidis and Lapata (2018) recently proposed a weakly supervised extension of the unsupervised ABAE. Their model, named Multi-seed Aspect Extractor, or MATE, initializes the aspect embedding A_k using the weighted average of the corresponding seed word embeddings (instead of the k-means centroids). To guarantee that the aspect embeddings will still be aligned with the K aspects of interest after training, Angelidis and Lapata (2018) keep the aspect and word embeddings fixed throughout training. In this work, we will show that the predictive power of seed words can be leveraged more effectively by considering each individual seed word as a more direct source of supervision during training.

[4]The reconstruction error can be efficiently estimated using contrastive max-margin objectives (Weston et al., 2011; Pennington et al., 2014).

2.2 Co-training

Co-training (Blum and Mitchell, 1998) is a classic multi-view learning method for semi-supervised learning. In co-training, classifiers over different feature spaces are encouraged to agree in their predictions on a large pool of unlabeled examples. Blum and Mitchell (1998) justify co-training in a setting where the different views are conditionally independent given the label. Several subsequent works have relaxed this assumption and shown co-training to be effective in much more general settings (Balcan et al., 2005; Chen et al., 2011; Collins and Singer, 1999; Clark et al., 2018). Co-training is also related to self-training (or bootstrapping) (Yarowsky, 1995), which trains a classifier using its own predictions and has been successfully applied for various NLP tasks (Collins and Singer, 1999; McClosky et al., 2006).

Recent research has successfully revisited these general ideas to solve NLP problems with modern deep learning methods. Clark et al. (2018) propose "cross-view training" for sequence modeling tasks by modifying Bi-LSTMs for *semi-supervised* learning. Ruder and Plank (2018) show that classic bootstrapping approaches such as tri-training (Zhou and Li, 2005) can be effectively integrated in neural networks for semi-supervised learning under domain shift. Our work provides further evidence that co-training can be effectively integrated into neural networks and combined with recent transfer learning approaches for NLP (Dai and Le, 2015; Howard and Ruder, 2018; Devlin et al., 2019; Radford et al., 2018), in a substantially different, *weakly supervised* setting where no ground-truth labels but only a few seed words are available for training.

Variable	Description
s	Segment (e.g., sentence) of a text review
K	Number of aspects of interest
D	Total number of seed words
$G_i \quad (i = 1; \ldots; K)$	Set of seed words for the i-th aspect
$h \in \mathbb{R}^d$	Segment embedding (student)
$c \in \mathbb{N}^D$	Bag-of-seed-words representation of s
$p = [p^1; \ldots; p^K]$	Student's aspect predictions
$q = [q^1; \ldots; q^K]$	Teacher's aspect predictions

Table 2: Notation.

2.3 Knowledge Distillation

Our approach is also related to the "knowledge distillation" framework (Buciluǎ et al., 2006; Ba and Caruana, 2014; Hinton et al., 2015), which has received considerable attention recently (Lopez-Paz et al., 2016; Kim and Rush, 2016; Furlanello et al., 2018; Wang, 2019). Traditional knowledge distillation aims at compressing a cumbersome model (teacher) to a simpler model (student) by training the student using both ground truth labels and the soft predictions of the teacher in a distillation objective. Our work also considers a student-teacher architecture and the distillation objective but under a considerably different, weakly supervised setting: (1) we do not use any labels for training and (2) we create conditions that allow the student to outperform the teacher; in turn, (3) we can use the student's predictions to learn a better teacher under co-training.

2.4 Problem Definition

Consider a corpus of text reviews from an entity domain (e.g., televisions, restaurants). Each review is split into segments (e.g., sentences, clauses). We also consider K pre-defined aspects of interest $(1; \ldots; K)$, including the "General" aspect, which we assume is the K-th aspect for simplicity. Different segments of the same review may be associated with different aspects but ground-truth aspect labels are *not* available for training. Instead, a small number of seed words G_k are provided for each aspect $k \in [K]$. Our goal is to use the corpus of training reviews and the available seed words $G = (G_1; \ldots; G_K)$ to train a classifier, which, given an unseen test segment s, predicts K aspect probabilities $p = [p^1; \ldots; p^K]$.

3 Our Student-Teacher Approach

We now describe our weakly supervised framework for aspect detection. We consider a student-teacher architecture (Figure 2), where the teacher

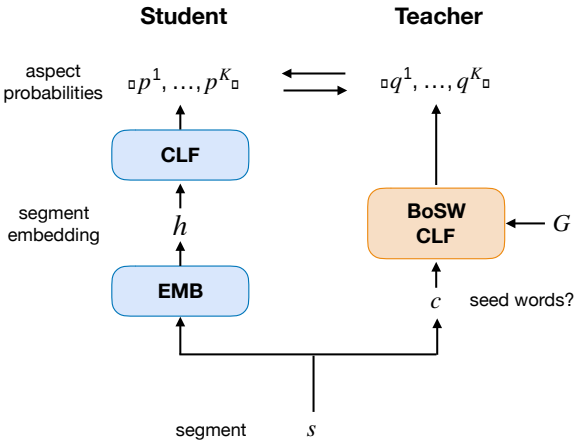

Figure 2: Our student-teacher approach for segment-level aspect detection using seed words.

is a bag-of-words classifier based solely on the provided seed words (i.e., a "bag-of-seed-words" classifier), and the student is an embedding-based neural network trained on data "softly" labeled by the teacher (as in the distillation objective). In the rest of this section, we describe the individual components of our student-teacher architecture and our proposed algorithm for performing updates.

3.1 Teacher: A Bag-of-Seed-Words Classifier

Our teacher model leverages the available seed words G that are predictive of the K aspects. Let D denote the total number of seed words in G. We can represent a segment s_i using a bag-of-seed-words representation $c_i \in \mathbb{N}^D$, where c_i^j encodes the number of times the j-th seed word occurs in s_i. (Note that c_i ignores the non-seed words.) The teacher's prediction for the k-th aspect is:

$$q_i^k = \frac{\exp\left(\sum_{j=1}^{D} \mathbb{1}\{j \in G_k\}\, c_i^j\right)}{\sum_{k'} \exp\left(\sum_{j=1}^{D} \mathbb{1}\{j \in G_{k'}\}\, c_i^j\right)}. \quad (1)$$

If no seed word appears in s, then the teacher predicts the "General" aspect by setting $q_i^K = 1$. Under this configuration the teacher uses seed words in a direct and intuitive way: it predicts aspect probabilities for the k-th aspect, which are proportional to the counts of the seed words under G_k, while if no seed word occurs in s, it predicts the "General" aspect. The classifier receives c_i as input and predicts $q_i = [q_i^1; \ldots; q_i^K]$.

Although the teacher only uses seed words to predict the aspect of a segment, we also expect

non-seed words to carry predictive power. Next, we describe the student network that learns to associate non-seed words with aspects.

3.2 Student: An Embedding-Based Network

Our student model is an embedding-based neural network: a segment is first embedded ($h_i = \mathrm{EMB}(s_i) \in \mathbb{R}^d$) and then classified to the K aspects ($p_i = \mathrm{CLF}(h_i)$) (see Section 2.1). The student does not use ground-truth aspect labels for training. Instead, it is trained by optimizing the distillation objective, i.e., the cross entropy between the teacher's (soft) predictions and the student's predictions:

$$H(q_i; p_i) = \sum_k q_i^k \log p_i^k \qquad (2)$$

While the teacher only uses the seed words in s_i to form its prediction q_i, the student uses all the words in s_i. Thus, using the distillation loss for training, the student learns to use both seed words and non-seed words to predict aspects. As a result, the student is able to generalize better than the teacher and *predict aspects even in segments that do not contain any seed words*. To regularize the student model, we apply L2 regularization to the classifier's weights and dropout regularization to the word embeddings (Srivastava et al., 2014). As we will show in Section 4, our student with this configuration outperforms the teacher in aspect prediction.

3.3 Iterative Co-Training

In this section, we describe our iterative co-training algorithm to cope with noisy seed words. The teacher in Section 3.1 considers each seed word equally, which can be problematic because not all seed words are equally good for predicting an aspect. In this work, we propose to estimate the predictive quality of each seed word in an unsupervised way. Our approach is inspired in the Model Bootstrapped Expectation Maximization (MBEM) algorithm of Khetan et al. (2018). MBEM is guaranteed to converge (under mild conditions) when the number of training data is sufficiently large and the worker quality is sufficiently high. Here, we treat seed words as "noisy annotators" and adopt an iterative estimation procedure similar to MBEM, as we describe next.

We model the predictive quality of the j-th seed word as a weight vector $z_j = \langle z_j^1, \ldots, z_j^K \rangle$, where

Algorithm 1 Iterative Seed Word Distillation

Input: $\{s_i\}_{i \in [N]}$, D seed words grouped into K disjoint sets $G = (G_1, \ldots, G_K)$
Output: $\hat{f}$: predictor function for segment-level aspect detection

Predict $\{q_i\}_{i \in [N]}$ (Eq. (1)) . *Apply teacher*
Repeat until convergence criterion
  Learn $\hat{f}$ (Eq. (2)) . *Train student*
  Predict $p_i = \hat{f}(s_i)$ $\{p_i\}_{i \in [N]}$. *Apply student*
  Update $\{z_j\}_{j \in [D]}$ (Eq. (4)) . *Update teacher*
  Predict $\{q_i\}_{i \in [N]}$ (Eq. (3)) . *Apply teacher*

z_j^k measures the strength of the association with the k-th aspect. We thus change the teacher to consider seed word quality. In particular, we replace Equation (1) by:

$$q_i^k = \frac{\exp \sum_{j=1}^{D} \mathbb{1}\{j \in G_k\} \hat{z}_j^k \, c_i^j}{\sum_{k'} \exp \sum_{j=1}^{D} \mathbb{1}\{j \in G_{k'}\} \hat{z}_j^{k'} \, c_i^j}; \qquad (3)$$

where $\hat{z}_j$ is the current estimate of z_j. As no ground-truth labels are available, we follow Khetan et al. (2018) and estimate z_j via Maximum Likelihood Estimation using the student's predictions as the current estimate of the ground truth labels. In particular, we assume that the prediction of the student for a training segment s_i is $t_i = \mathrm{argmax}_k \, p_i^k$. Then, for each seed word we compute the quality estimate for the k-th aspect using the student's predictions for N segments:

$$\hat{z}_j^k = \frac{\sum_{i=1}^{N} \mathbb{1}\{c_i^j > 0\} \mathbb{1}\{t_i = k\}}{\sum_{k'} \sum_{i=1}^{N} \mathbb{1}\{c_i^j > 0\} \mathbb{1}\{t_i = k'\}}. \qquad (4)$$

According to Equation (4), the quality of the j-th seed word is estimated according to the student-teacher agreement on segments where the seed word appears.

Building upon the previous ideas, we present our Iterative Seed Word Distillation (ISWD) algorithm for effectively leveraging the seed words for fine-grained aspect detection. Each round of ISWD consists of the following steps (Algorithm 1): (1) we apply the teacher on unlabeled training segments to get predictions q_i (without considering seed word qualities); (2) we train the student using the teacher's predictions in the distillation

objective of Equation (2);[5] (3) we apply the student in the training data to get predictions p_i; and (4) we update the seed word quality parameters using the student's predictions in Equation (4).

In contrast to MATE, which uses the validation set (with aspect labels) to estimate seed weights in an initialization step, our proposed method is an unsupervised approach to modeling and adapting the seed word quality during training. We stop this iterative procedure after the disagreement between the student's and teacher's hard predictions in the training data stops decreasing. We empirically observe that 2-3 rounds are sufficient to satisfy this criterion. This observation also agrees with Khetan et al. (2018), who only run their algorithm for two rounds.

4 Experiments

We evaluate our approach to aspect detection on several datasets of product and restaurant reviews.

4.1 Experimental Settings

Datasets. We train and evaluate our models on Amazon product reviews for six domains (Laptop Bags, Keyboards, Boots, Bluetooth Headsets, Televisions, and Vacuums) from the OPOSUM dataset (Angelidis and Lapata, 2018), and on restaurant reviews in six languages (English, Spanish, French, Russian, Dutch, Turkish) from the SemEval-2016 Aspect-based Sentiment Analysis task (Pontiki et al., 2016). Aspect labels (9-class for product reviews and 12-class for restaurant reviews) are available for each segment[6] of the validation and test sets. The restaurant reviews also come with training aspect labels, which we only use for training the fully supervised models. For a fair comparison, we use exactly the same 30 seed words (per aspect and domain) used in Angelidis and Lapata (2018) for the product reviews and use the same extraction method described in Angelidis and Lapata (2018) to extract 30 seed words for the restaurant reviews. See Appendix A for more dataset details.

[5]Note that the quality-aware loss function proposed in Khetan et al. (2018), which is an alternative form of noise-aware loss functions (Natarajan et al., 2013), is equivalent to our distillation loss: using the log loss as $l(.)$ in Equation (4) of Khetan et al. (2018) yields the cross entropy loss.

[6]In product reviews, elementary discourse units (EDUs) are used as segments. In restaurant reviews, sentences are used as segments.

Experimental Procedure. For a fair comparison, we use exactly the same pre-processing (tokenization, stemming, and word embedding) and evaluation procedure as in Angelidis and Lapata (2018). For each domain, we train our model on the training set without using any aspect labels, and only use the seed words G via the teacher. For each model, we report the average test performance over 5 different runs with the parameter configuration that achieves best validation performance. As evaluation metric, we use the micro-averaged F1.

Model Configuration. For the student network, we experiment with various modeling choices for segment representations: bag-of-words (BOW) classifiers, the unweighted average of word2vec embeddings (W2V), the weighted average of word2vec embeddings using bilinear attention (Luong et al., 2015) (same setting as He et al. (2017); Angelidis and Lapata (2018)), and the average of contextualized word representations obtained from the second-to-last layer of the pre-trained (self-attention based) BERT model (Devlin et al., 2019), which uses multiple self-attention layers (Vaswani et al., 2017) and has been shown to achieve state-of-the-art performance in many downstream NLP applications. For the English product reviews, we use the base uncased BERT model. For the multilingual restaurant reviews, we use the multilingual cased BERT model.[7]

In iterative co-training, we train the student network to convergence in each iteration (which may require more than one epoch over the training data). Moreover, we observed that the iterative process is more stable when we interpolate between weights of the previous iteration and the estimated updates instead of directly applying the estimated seed weight updates (according to Equation (3)).

Model Comparison. For a robust evaluation of our approach, we compare the following models and baselines:

LDA-Anchors: The topic model of Lund et al. (2017) using seed words as "anchors."

ABAE: The unsupervised autoencoder of He et al. (2017), where the learned topics were

[7]Both models can be found in https://github.com/google-research/bert/blob/master/multilingual.md. The multilingual cased BERT model is recommended by the authors instead of the multilingual uncased BERT model.

Method	Product Review Domain						AVG
	Bags	Keyboards	Boots	Headsets	TVs	Vacuums	
LDA-Anchors (Lund et al., 2017)	33.5	34.7	31.7	38.4	29.8	30.1	33.0
ABAE (He et al., 2017)	38.1	38.6	35.2	37.6	39.5	38.1	37.9
MATE (Angelidis and Lapata, 2018)	46.2	43.5	45.6	52.2	48.8	42.3	46.4
MATE-unweighted	41.6	41.3	41.2	48.5	45.7	40.6	43.2
MATE-MT (best performing)	48.6	45.3	46.4	54.5	51.8	47.7	49.1
Teacher	55.1	52.0	44.5	50.1	56.8	54.5	52.2
Student-BoW	57.3	56.2	48.8	59.8	59.6	55.8	56.3
Student-W2V	59.3	57.0	48.3	**66.8**	**64.0**	57.0	58.7
Student-W2V-RSW	51.3	57.2	46.6	63.0	62.1	57.1	56.2
Student-ATT	60.1	55.6	49.9	66.6	63.4	58.2	58.9
Student-BERT	**61.4**	**57.5**	**52.0**	66.5	63.0	**60.4**	**60.2**

Table 3: Micro-averaged F1 reported for 9-class EDU-level aspect detection in product reviews.

Method	Restaurant Review Language						AVG
	En	Sp	Fr	Ru	Du	Tur	
W2V-Gold	58.8	50.4	50.4	69.3	51.4	55.7	56.0
BERT-Gold	63.1	51.6	50.5	64.6	53.5	55.3	56.4
MATE	41.0	24.9	17.8	18.4	36.1	39.0	29.5
MATE-unweighted	40.3	18.3	19.2	21.8	31.5	25.2	26.1
Teacher	44.9	41.8	34.1	54.4	40.7	30.2	41.0
Student-W2V	47.2	40.9	32.4	**59.0**	42.1	42.3	44.0
Student-ATT	47.8	41.7	32.9	57.3	44.1	45.5	44.9
Student-BERT	**51.8**	**42.0**	**39.2**	58.0	**43.0**	**45.0**	**46.5**

Table 4: Micro-averaged F1 reported for 12-class sentence-level aspect detection in restaurant reviews. The fully supervised *-Gold models are not directly comparable with the weakly supervised models.

manually mapped to aspects.

MATE-*: The MATE model of Angelidis and Lapata (2018) with various configurations: initialization of the aspect embeddings A_k using the unweighted/weighted average of seed word embeddings and an extra multi-task training objective (MT).[8]

Teacher: Our bag-of-seed-words teacher.

Student-*: Our student network trained with various configurations for the EMB function.

***-Gold:** Supervised models trained using ground truth aspect labels, which are only available for restaurant reviews. These models are not directly comparable with the other models and baselines.

4.2 Experimental Results

Tables 3 and 4 show the results for aspect detection on product and restaurant reviews, respectively. The rightmost column of each table reports the average performance across the 6 domains/languages.

MATE-* models outperform ABAE. Using the seed words to initialize aspect embeddings leads to more accurate aspect predictions than mapping the learned (unsupervised) topics to aspects.

LDA-Anchors performs worse than MATE-* models. Although averages of seed words were used as "anchors" in the "Tandem Anchoring" algorithm, we observed that the learned topics did not correspond to our aspects of interest.

The teacher effectively leverages seed words. By leveraging the seed words in a more direct way, Teacher is able to outperform the MATE-* models. Thus, we can use Teacher's predictions as supervision for the student, as we describe next.

The student outperforms the teacher. Student-BoW outperforms Teacher: the two models have the same architecture but Teacher only considers seed words; regularizing Student's weights en-

[8]The multi-task training objective in MATE requires datasets from different domains but same language, thus it cannot be applied in our datasets of restaurant reviews.

courages Student to mimic the noisy aspect predictions of Teacher by also considering non-seed words for aspect detection. The benefits of our distillation approach are highlighted using neural networks with word embeddings. Student-W2V outperforms both Teacher and Student-BoW, showing that obtaining segment representations as the average of word embeddings is more effective than using bag-of-words representations for this task.

The student outperforms previous weakly supervised models even in one co-training round. Student-ATT outperforms MATE-unweighted (by 36.3% in product reviews and by 52.2% in restaurant reviews) even in a single co-training round: although the two models use exactly the same seed words (without weights), pre-trained word embeddings, EMB function, and CLF function, our student-teacher approach leverages the available seed words more effectively as noisy supervision than just for initialization. Also, using our approach, we can explore more powerful methods for segment embedding without the constraint of a fixed word embedding space. Indeed, using contextualized word representations in Student-BERT leads to the best performance over all models.

As expected, our weakly supervised approach does not outperform the fully supervised (*-Gold) models. However, our approach substantially reduces the performance gap between weakly supervised approaches and fully supervised approaches by 62%. The benefits of our student-teacher approach are consistent across all datasets, highlighting the predictive power of seed words across different domains and languages.

The student leverages non-seed words. To better understand the extent to which non-seed words can predict the aspects of interest, we experiment with completely removing the seed words from Student-W2V's input during training (Student-W2V-RSW method; see Figure 3). Thus, in this setting, Student-W2V-RSW is forced to only use non-seed words to detect aspects. Note that the co-training assumption of conditionally independent views (Blum and Mitchell, 1998) is satisfied in this setting, where Teacher is only using seed words and Student-W2V is only using non-seed words. Student-W2V-RSW effectively learns to use non-seed words to predict aspects and performs better than Teacher (but worse than Student-W2V, which considers both seed and non-seed words). For additional ablation experiments, see Appendix A.

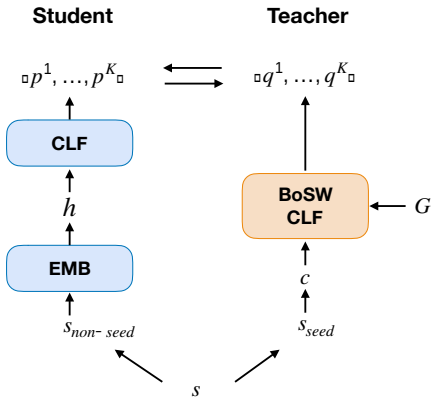

Figure 3: Our weakly supervised co-training approach when seed words are removed from the student's input (RSW baseline). Segment $s_{non-seed}$ is an edited version of s, where we replace each seed word in s by an "UNK" special token (like out-of-vocabulary words).

Method	Initial	Iterative
Product Reviews (AVG)		
MATE	46.4	-
Teacher / Student-W2V	52.2 / 58.7	**58.5 / 59.7**
Teacher / Student-BERT	52.2 / 60.2	**58.6 / 60.8**
Restaurant Reviews (En)		
MATE	29.5	-
Teacher / Student-W2V	44.9 / 47.2	**45.8 / 49.0**
Teacher / Student-BERT	44.9 / 51.8	**49.8 / 53.4**

Table 5: Micro-averaged F1 scores during the first round (middle column) and after iterative co-training (right column) in product reviews (top) and restaurant reviews (bottom).

Iterative co-training copes with noisy words. Further performance improvement in Teacher and Student-* can be observed with the iterative co-training procedure of Section 3.3. Table 5 reports the performance of Teacher and Student-* after co-training for both product reviews (top) and English restaurant reviews (bottom). (For more detailed, per-domain results, see Appendix A.) Compared to the initial version of Teacher that does not model the quality of the seed words, iterative co-training leads to estimates of seed word quality that improve Teacher's performance up to 12.3% (in product reviews using Student-BERT).

A better teacher leads to a better student. Co-training leads to improved student performance in both datasets (Table 5). Compared to MATE, which uses the validation set to estimate the seed weights as a pre-processing step, we estimate

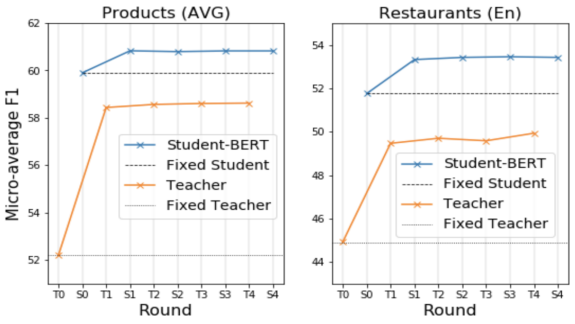

Figure 4: Co-training performance for each round reported for product reviews (left) and restaurant reviews (right). T<*i*>and S<*i*>correspond to the teacher's and student's performance, respectively, at the *i*-th round.

and iteratively adapt the seed weights using the student-teacher disagreement, which substantially improves performance. Across the 12 datasets, Student-BERT leads to an average absolute increase of 14.1 F1 points.

Figure 4 plots Teacher's and Student-BERT's performance after each round of co-training. Most of the improvement for both Teacher and Student-BERT is gained in the first two rounds of co-training: "T0" (in Figure 4) is the initial teacher, while "T1" is the teacher with estimates of seed word qualities, which leads to more accurate predictions, e.g., in segments with multiple seed words from different aspects.

5 Conclusions and Future Work

We presented a weakly supervised approach for leveraging a small number of seed words (instead of ground truth aspect labels) for segment-level aspect detection. Our student-teacher approach leverages seed words more directly and effectively than previous weakly supervised approaches. The teacher model provides weak supervision to a student model, which generalizes better than the teacher by also considering non-seed words and by using pre-trained word embeddings. We further show that iterative co-training lets us estimate the quality of the (possibly noisy) seed words. This leads to a better teacher and, in turn, a better student. Our proposed method consistently outperforms previous weakly supervised methods in 12 datasets, allowing for seed words from various domains and languages to be leveraged for aspect detection. Our student-teacher approach could be applied for any classification task for which a small set of seed words describe each class. In future work, we plan to extend our framework to multi-task settings, and to incorporate interaction to learn better seed words.

Acknowledgments

We thank the anonymous reviewers for their constructive feedback. This material is based upon work supported by the National Science Foundation under Grant No. IIS-15-63785.

References

Stefanos Angelidis and Mirella Lapata. 2018. Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Sanjeev Arora, Rong Ge, Yoni Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. 2013. A practical algorithm for topic modeling with provable guarantees. In *Proceedings of the International Conference on International Conference on Machine Learning*.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *Proceedings of the International Conference on Learning Representations*.

Jimmy Ba and Rich Caruana. 2014. Do deep nets really need to be deep? In *Advances in Neural Information Processing Systems*.

Maria-Florina Balcan, Avrim Blum, and Ke Yang. 2005. Co-training and expansion: Towards bridging theory and practice. In *Advances in Neural Information Processing Systems*.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the Annual Conference on Computational Learning Theory*.

Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Minmin Chen, Kilian Q Weinberger, and John Blitzer. 2011. Co-training for domain adaptation. In *Advances in Neural Information Processing Systems*.

Kevin Clark, Minh-Thang Luong, Christopher D Manning, and Quoc Le. 2018. Semi-supervised sequence modeling with cross-view training. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Michael Collins and Yoram Singer. 1999. Unsupervised models for named entity classification. In *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

Andrew M Dai and Quoc V Le. 2015. Semi-supervised sequence learning. In *Advances in Neural Information Processing Systems*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Vanessa Wei Feng and Graeme Hirst. 2012. Text-level discourse parsing with rich linguistic features. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. 2018. Born-again neural networks. In *International Conference on Machine Learning*.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2017. An unsupervised neural attention model for aspect extraction. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Mohit Iyyer, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber, and Hal Daumé III. 2016. Feuding 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*.

Giannis Karamanolakis, Daniel Hsu, and Luis Gravano. 2019. Training neural networks for aspect extraction using descriptive keywords only. In *Proceedings of the Second Learning from Limited Labeled Data Workshop*.

Ashish Khetan, Zachary C Lipton, and Anima Anandkumar. 2018. Learning from noisy singly-labeled data. In *Proceedings of the International Conference on Learning Representations*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*.

Yoon Kim and Alexander M Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Pengfei Liu, Shafiq Joty, and Helen Meng. 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*.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. 2016. Unifying distillation and privileged information. In *Proceedings of the International Conference on Learning Representations*.

Bin Lu, Myle Ott, Claire Cardie, and Benjamin K Tsou. 2011. Multi-aspect sentiment analysis with topic models. In *2011 IEEE International Conference on Data Mining Workshops*. IEEE.

Jeffrey Lund, Connor Cook, Kevin Seppi, and Jordan Boyd-Graber. 2017. Tandem anchoring: A multiword anchor approach for interactive topic modeling. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

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*.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*.

David McClosky, Eugene Charniak, and Mark Johnson. 2006. Effective self-training for parsing. In *Proceedings of the 2006 Conference of the North American Chapter of the Association of Computational Linguistics*. Association for Computational Linguistics.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. 2013. Learning with noisy labels. In *Advances in Neural Information Processing Systems*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, AL-Smadi Mohammad, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*.

Soujanya Poria, Erik Cambria, and Alexander Gelbukh. 2016. Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems*, 108:42–49.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *https://blog.openai.com/language-unsupervised*.

Sebastian Ruder, Parsa Ghaffari, and John G Breslin. 2016. A hierarchical model of reviews for aspect-based sentiment analysis. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*.

Sebastian Ruder and Barbara Plank. 2018. Strong baselines for neural semi-supervised learning under domain shift. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Akash Srivastava and Charles Sutton. 2071. Autoencoding variational inference for topic models. In *Proceedings of the International Conference on Learning Representations*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

Weiran Wang. 2019. Everything old is new again: A multi-view learning approach to learning using privileged information and distillation. *arXiv preprint arXiv:1903.03694*.

Jason Weston, Samy Bengio, and Nicolas Usunier. 2011. Wsabie: Scaling up to large vocabulary image annotation. In *Twenty-Second International Joint Conference on Artificial Intelligence*.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Towards universal paraphrastic sentence embeddings. *In Proceedings of the International Conference on Learning Representations*.

John Wieting and Kevin Gimpel. 2017. Revisiting recurrent networks for paraphrastic sentence embeddings. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Annual Meeting of the Association for Computational Linguistics*.

Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, page e1253.

Zhi-Hua Zhou and Ming Li. 2005. Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Transactions on Knowledge & Data Engineering*, (11):1529–1541.

A Appendix

For reproducibility, we provide more information on datasets (Section A.1) and implementation details (Section A.2), and report more detailed evaluation results (Section A.3).

A.1 Datasets

In this section, we describe all details of the datasets of product and restaurant reviews, and report dataset statistics.

Product Reviews. The OPOSUM dataset (Angelidis and Lapata, 2018) is a subset of the Amazon Product Dataset (McAuley et al., 2015), which contains Amazon reviews from 6 domains: Laptop Bags, Keyboards, Boots, Bluetooth Headsets, Televisions, and Vacuums. The validation and test segments of each domain have been manually annotated with 9 aspects (Table 4). The reviews of each domain are already segmented by Angelidis and Lapata (2018) into elementary discourse units (EDUs) using a Rhetorical Structure Theory parser (Feng and Hirst, 2012). The average number of training, validation, and test segments across domains is around 1 million, 700, and 700 segments, respectively. Segment statistics per domain are reported in the supplementary material of Angelidis and Lapata (2018).

Restaurant Reviews. The datasets used in the SemEval-2016 Aspect-based Sentiment Analysis task (Pontiki et al., 2016) contain reviews for multiple domains and languages. Here, we use the six corpora of multilingual (English, Spanish, French, Russian, Dutch, Turkish) restaurant reviews. The training, validation, and test segments have been manually annotated with 12 aspects, which are shared across languages:

1. Restaurant#General
2. Food#Quality
3. Service#General
4. Ambience#General
5. Food#Style_Options
6. Food#Prices
7. Restaurant#Miscellaneous
8. Restaurant#Prices
9. Drinks#Quality
10. Drinks#Style_Options
11. Location#General
12. Drinks#Prices

The reviews of each language are already segmented into sentences. The average number of training and test segments across languages is around 2500 and 800 segments respectively. The training segments of restaurant reviews are significantly fewer than the training segments of product reviews. Therefore, for non-English reviews we report results after a single co-training round. For our co-training experiments we augment the English reviews dataset with 50,000 English reviews randomly sampled from the Yelp Challenge corpus.[9]

A.2 Implementation Details

For a fair comparison, for the product reviews we use the 200-dimensional word2vec embeddings provided by Angelidis and Lapata (2018) and the base uncased BERT model.[10] For the restaurant reviews, we use the 300-dimensional multilingual word2vec embeddings provided by Ruder et al. (2016) and the multilingual cased BERT model.[11] The student's parameters are optimized using Adam (Kingma and Ba, 2014) with learning rate 0.005 and mini-batch size 50. After each co-training round we divide the learning rate by 10. We apply dropout in the word embeddings and the last hidden layers of the classifiers (Srivastava et al., 2014) with rate 0.5.

A.3 More Results

Table 5 reports detailed per-domain results. "Teacher (symmetric)" is a simpler version of Teacher that randomly guesses the aspect of segments with no seed words. For Student-W2V we report additional ablation experiments. The *-ISWD models correspond to student or teacher models after multiple rounds of co-training until convergence.

[9] https://www.yelp.com/dataset/challenge

[10] https://github.com/google-research/bert#pre-trained-models

[11] https://github.com/google-research/bert/blob/master/multilingual.md

Bags	Keyboards	Boots	Headsets	TVs	Vacuums
Size/Fit	Feel/Comfort	Comfort	Sound	Image	Accessories
Quality	Layout	Size	Comfort	Sound	Ease of Use
Looks	Build Quality	Look	Ease of Use	Connectivity	Suction Power
Compartments	Extra Function.	Materials	Connectivity	Customer Serv.	Build Quality
Handles	Connectivity	Durability	Durability	Ease of Use	Noise
Protection	Price	Weather Resist.	Battery	Price	Weight
Price	Noise	Price	Price	Apps/Interface	Customer Serv.
Customer Serv.	Looks	Color	Look	Size/Look	Price
General	General	General	General	General	General

Table 4: The 9 aspect classes per domain of product reviews (OPOSUM).

Method	Product Review Domain						AVG
	Bags	Keyboards	Boots	Headsets	TVs	Vacuums	
Previous Approaches							
LDA-Anchors (Lund et al., 2017)	33.5	34.7	31.7	38.4	29.8	30.1	33.0
ABAE (He et al., 2017)	38.1	38.6	35.2	37.6	39.5	38.1	37.9
MATE (Angelidis and Lapata, 2018)	46.2	43.5	45.6	52.2	48.8	42.3	46.4
MATE-unweighted	41.6	41.3	41.2	48.5	45.7	40.6	43.2
MATE-MT (best performing)	48.6	45.3	46.4	54.5	51.8	47.7	49.1
Our Approach: Single Round Co-training							
Teacher (symmetric)	38.9	27.7	30.3	34.0	33.5	35.6	33.3
Teacher	55.1	52.0	44.5	50.1	56.8	54.5	52.2
Student-BoW	57.3	56.2	48.8	59.8	59.6	55.8	56.3
Student-W2V	59.3	57.0	48.3	**66.8**	**64.0**	57.0	58.7
Student-W2V-RSW	51.3	57.2	46.6	63.0	62.1	57.1	56.2
Student-W2V w/o L2 Reg	56.3	56.6	48.8	59.8	58.4	54.7	55.7
Student-W2V w/o dropout	56.4	56.2	48.1	59.4	57.4	54.2	55.3
Student-W2V w/o emb fine-tuning	58.7	53.6	42.8	62.2	56.3	54.3	54.6
Student-W2V w/o soft targets	57.2	57.4	47.1	61.7	58.3	55.0	56.1
Student-ATT	60.1	55.6	49.9	66.6	63.4	58.2	58.9
Student-BERT	**61.4**	**57.5**	**52.0**	66.5	63.0	**60.4**	**60.2**
Our Approach: Iterative Co-training							
Teacher-ISWD (St: W2V)	59.3	58.2	50.6	63.6	61.0	58.4	58.5
Teacher-ISWD (St: ATT)	59.6	58.0	50.6	62.4	60.6	59.0	58.3
Teacher-ISWD (St: BERT)	57.7	59.6	50.4	64.0	60.9	59.1	58.6
Student-W2V-ISWD	58.7	57.0	52.6	67.6	63.2	58.8	59.7
Student-ATT-ISWD	59.6	55.9	51.0	**67.9**	65.6	59.8	60.0
Student-BERT-ISWD	59.1	**59.0**	**53.9**	65.8	**66.1**	**61.0**	**60.8**

Table 5: Micro-averaged F1 reported for 9-class EDU-level aspect detection in product reviews.