DILBERT: Customized Pre-Training for Domain Adaptation with Category Shift, with an Application to Aspect Extraction

Entony Lekhtman [1] **Yftah Ziser**[2] **Roi Reichart**[3]

[1]Faculty of Computer Science, Technion, IIT
[2]School of Informatics, University of Edinburgh
[3]Faculty of Industrial Engineering and Management, Technion, IIT

`tony1@campus.technion.ac.il yftah.ziser@ed.ac.uk roiri@technion.ac.il`

Abstract

The rise of pre-trained language models has yielded substantial progress in the vast majority of Natural Language Processing (NLP) tasks. However, a generic approach towards the pre-training procedure can naturally be sub-optimal in some cases. Particularly, fine-tuning a pre-trained language model on a source domain and then applying it to a different target domain, results in a sharp performance decline of the eventual classifier for many source-target domain pairs. Moreover, in some NLP tasks, the output categories substantially differ between domains, making adaptation even more challenging. This, for example, happens in the task of aspect extraction, where the aspects of interest of reviews of, e.g., restaurants or electronic devices may be very different. This paper presents a new fine-tuning scheme for BERT, which aims to address the above challenges. We name this scheme *DILBERT: Domain Invariant Learning with BERT*, and customize it for aspect extraction in the unsupervised domain adaptation setting. DILBERT harnesses the categorical information of both the source and the target domains to guide the pre-training process towards a more domain and category invariant representation, thus closing the gap between the domains. We show that DILBERT yields substantial improvements over state-of-the-art baselines while using a fraction of the unlabeled data, particularly in more challenging domain adaptation setups.[1]

1 Introduction

Aspect-based sentiment analysis (ABSA) (Thet et al., 2010), extracting aspect-sentiment pairs for products or services from reviews, is a widely researched task in both academia and industry. ABSA allows a fine-grained and realistic evaluation of reviews, as real-world reviews typically

Figure 1: A review from the Restaurants domain. We use BIO tags to mark the spans of the aspects. Each aspect belongs to one category: For example, the *waiter* aspect belongs to the *service* category. In this review, the word *food* serves as both an aspect and a category.

do not convey a homogeneous sentiment but rather communicate different sentiments for different aspects of the reviewed item or service. For example, while the overall sentiment of the review in Figure 1 is unclear, the sentiment towards the service, food, and location of the restaurant is very decisive. Moreover, even when the overall sentiment of the review is clear, ABSA provides more nuanced and complete information about its content.

In this paper, we focus on the sub-task of aspect extraction (AE, a.k.a opinion targets extraction): Extracting from opinionated texts the aspects on which the reader conveys sentiment. For example, in Figure 1, the waiter, food, and the views of the city are aspects derived from broader categories: service, food, and location. This task is characterized by a multiplicity of domains, as reviews and other opinionated texts can be written about a variety of products, services as well as many other issues. Moreover, the aspect categories of interest often differ between these domains.

As for most NLP tasks and applications, AE research has recently made substantial progress. While Transformer (Vaswani et al., 2017) based pre-trained models (Devlin et al., 2019; Liu et al., 2019) have pushed results substantially forward, they rely on in-domain labeled data to achieve their strong results. Annotating such data for multiple domains is costly and laborious, which is one of the major bottlenecks for developing and deploying NLP systems. As noted above, AE forms a particu-

---

larly challenging variant of the domain adaptation problem, as the aspect categories of interest tend to change across domains.

A well-established approach for addressing the above bottleneck is Domain Adaptation (DA) (Blitzer et al., 2006; Ben-David et al., 2007). DA, training models on source domain labeled data so that they can effectively generalize to different target domains, is a long-standing research challenge. While the target domain labeled data availability in DA setups ranges from little (supervised DA (Daumé III, 2007)) to none (unsupervised DA (Ramponi and Plank, 2020)), unlabeled data is typically available in both source and target domains. This paper focuses on unsupervised DA as we believe it is a more realistic and practical scenario.

Due to the great success of deep learning models, DA through representation learning (Blitzer et al., 2007; Ziser and Reichart, 2017), i.e., learning a shared representation for both the source and target domains, has recently become prominent (Ziser and Reichart, 2018a; Ben-David et al., 2020). Of particular importance to this line of work are approaches that utilize pivot features (Blitzer et al., 2006) that: (a) frequently appear in both domains; and (b) have high mutual information (MI) with the task label. While pivot-based methods achieve state-of-the-art results in many text classification tasks (Ziser and Reichart, 2018a; Miller, 2019; Ben-David et al., 2020), it is not trivial to successfully apply them on tasks such as AE. This stems from two reasons: First, AE is a sequence tagging task with multiple labels for each input example (i.e., word-level labels for input sentences). For a feature to meet the second condition for being a pivot (high MI with the task label), further refinement of the pivot definition is required. Second, different domains often differ in their aspect categories and hence if a feature is highly correlated with a source domain label (aspect category), this is not indicative of its being correlated with the (different) aspect categories of the target domain.

To overcome these limitations, we present *DILBERT: Domain Invariant Learning with BERT*, a customized fine-tuning procedure for AE in an unsupervised DA setup. More specifically, DILBERT employs a variant of the BERT masked language modeling (MLM) task such that hidden tokens are chosen by their semantic similarity to the categories rather than randomly. Further, it employs a new pre-training task: The prediction of which categories appear in the input text. Notice that unlabeled text does not contain supervision for our pre-training tasks and we hence have to use distant supervision as an approximation.

In our unsupervised DA experiments we consider laptop and restaurant reviews, where for the restaurant domain we consider two variants, that differ in their difficulty. Our best performing model outperforms the strongest baseline by over 5% on average and over 13% on the most challenging setup, while using only a small fraction of the unlabeled data. Moreover, we show that our pre-training procedure is very effective in resource-poor setups, where unlabeled data is scarce.

2 Background and Previous Work

While both aspect extraction and domain adaptation are active fields of research, research on their intersection is not as frequent. We hence first describe in-domain AE, and then continue with a survey of DA, focusing on pivot-based unsupervised DA. Finally, we describe works at the intersection of both problems.

2.1 Aspect Extraction

Early in-domain aspect extraction works heavily rely on feature engineering, often feeding graphical models with linguistic-driven features such as part-of-speech (POS) tags (Jin et al., 2009), WordNet attributes and word frequencies (Li et al., 2010). The SemEval ABSA task releases (Pontiki et al., 2014, 2015, 2016) and the rise of deep learning have pushed AE research substantially forward.

Liu et al. (2015) applied Recurrent Neural Networks (RNN) combined with simple linguistic features to outperform feature-rich Conditional Random Field (CRF) models. Wang et al. (2016) showed that stacking a CRF on top of an RNN further improves the results. Recently, Xu et al. (2019) tuned BERT for AE and obtained additional improvements, demonstrating the effectiveness of massive pre-training with unlabeled data for this task. Tian et al. (2020) proposed new pre-training tasks based on automatically extracted sentiment words and aspect terms. Finally, Jiang et al. (2019) released the Multi-Aspect Multi-Sentiment (MAMS) dataset, where each sentence contains at least two different aspects with different sentiment polarities, making it more challenging compared to the SemEval datasets.

2.2 Domain Adaptation

DA is a fundamental challenge in machine learning in general and NLP in particular. In this work, we focus on unsupervised DA, in which labeled data is available from the source domain and unlabeled data is available from both the source and the target domains. DA approaches include instance re-weighting (Mansour et al., 2008; Gong et al., 2020), sub-sampling from both domains (Chen et al., 2011) and learning a shared representation for the source and target domains (Blitzer et al., 2007; Ganin et al., 2016; Ziser and Reichart, 2018b). This section focuses on the latter approach which has become prominent due to the success of deep learning. Indeed, our proposed method, as well as the previous methods and the baselines we compare to, follow this approach.

Unsupervised DA via Shared Representation
The shared representation approach to unsupervised DA typically consists of two major steps: (a) A representation model is trained using the unlabeled data from the source and target domains; and (b) A task-specific classifier is stacked on top of the representation model of step (a) and fine-tuned using the source domain labeled data. The fine-tuned model is then applied to the target domain test data, hoping that the domain-invariant feature space would mitigate the domain gap. Many works have followed this avenue (e.g., (Chen et al., 2012; Louizos et al., 2016; Ganin et al., 2016)) and a comprehensive survey is beyond the scope of this paper. Below we discuss the line of work which is most relevant to our model as well as to most previous unsupervised DA for AE work.

Shared Representation Using Pivot Features
Pivots features, proposed by Blitzer et al. (2006, 2007) through their structural correspondence learning (SCL) framework, are features that meet both the domain frequency and the source-domain task label correlation conditions, defined in §1. The authors use the distinction between pivot and non-pivot features (features that do not meet at least one of the criteria, as long as they are frequent in one of the domains) in order to learn a shared cross-domain representation model. The main idea is to utilize the pivot features to extract cross-domain and task-relevant information from non-pivot features, which is done through non-pivot to pivot feature mapping. This way the induced representation consists of the cross-domain and task-relevant

information of both feature types.

Blitzer et al. (2006, 2007) learned linear non-pivot to pivot mappings that do not exploit the input structure (e.g., the structure of the review document in a sentiment classification task). A series of consecutive works alleviated these limitations. For example, Ziser and Reichart (2017) used a feed-forward neural network to learn the mapping, also exploiting the semantic similarity between pivots. Later (Ziser and Reichart, 2018b, 2019) proposed PBLM, a pivot-based language model, which also exploits the structure of the input text . Recently, Ben-David et al. (2020) integrated these ideas into the BERT architecture. They changed its Masked Language Modeling task so that pivots are more often masked than non-pivots, and the model should predict if a token is a pivot or not, and then identify the token only in the former case.

Despite the great success of this line of work on unsupervised DA for sentiment classification, the adaptation of the proposed ideas to sequence tagging tasks with cross-domain category shift is challenging. In this paper we solve this challenge and demonstrate that our solution yields state-of-the-art results on unsupervised DA for AE. We next survey existing work on this problem.

2.3 Domain Adaptation for AE

Like most recent DA works, the shared representation approach is prominent in DA for AE, where previous works align the different domains using syntactic patterns, reasoning that such patterns are robust across domains. For example, Ding et al. (2017) learn a shared representation by training an RNN to predict rule-based syntactic patterns, populated by a pre-defined sentiment lexicon. While dependency relations-based rules are known to improve in-domain aspect extraction (Qiu et al., 2011; Wang et al., 2016), using hand-crafted patterns with a pre-defined sentiment lexicon heavily relies on prior knowledge and might not be robust when adapting to new, more challenging domains.

Similarly, inspired by the pivot-based modeling approach of Blitzer et al. (2006), Wang and Pan (2018) train a recursive recurrent network to predict source and target dependency trees (obtained by an off-the-shelf parser (Klein and Manning, 2003)). Then, they jointly train the model to predict aspect and opinion words. Likewise, Pereg et al. (2020) incorporate syntactic knowledge from an external parser (Dozat and Manning, 2017) into BERT via

its self-attention mechanism. Relying on supervised parsers, these approaches naturally suffer from the degradation of such parsers when applied to resource-poor domains (e.g., user-generated content) or languages. Moreover, the work of Wang and Pan (2018) requires additional human annotation for opinion word labels, which might not be available for new domains.

Li et al. (2019) avoids the need for external resources (except from an opinion lexicon), by applying a dual memory mechanism combined with a gradient reversal layer (Ganin et al., 2016), in a model that jointly learns to predict aspect and opinion terms. Recently, Gong et al. (2020) presented the Unified Domain Adaptation (UDA) approach, the first to apply a pre-trained language encoder (BERT) to our task. Particularly, they apply self-supervised POS and dependency relation information as an auxiliary training task in order to bias BERT towards domain-invariant representations. Then, they apply instance re-weighting, and this way they perform DA at both the representation and the training instance level.

While syntactic pivot-based models contributed to unsupervised DA for AE, they do not harness the semantic properties of the involved domains. Moreover, some previous work rely on external, resource-intensive syntactic models and on manual rules. Finally, the only previous work that applies a pre-trained language encoder (BERT) also focuses on syntax-driven adaptation. Our approach exploits the power of BERT for learning a cross-domain shared representation, but with semantically-driven, self-supervised pre-training tasks.

3 Domain Adaptation with DILBERT

In this section, we introduce DILBERT, a new fine-tuning scheme for BERT (Devlin et al., 2019), customized for domain and category shift. Recall that BERT performs two pre-training tasks: (a) Masked Language Modeling (MLM), where some of the input tokens are randomly masked and the model should predict them based on their context; and (b) Next Sentence Prediction (NSP), where the model is provided with sentence pairs from its training data and it should predict whether one sentence is indeed followed by the other.

DILBERT modifies the MLM task and presents a new pre-training task. Notice that DILBERT is applied to a BERT model that has already been trained on general text – text that is not directly related to the source and target domains of interest – with the standard MLM and NSP tasks. Hence, DILBERT can be seen as a fine-tuning step on the unlabeled data from the source and target domains. After DILBERT is applied, a task classifier is added on top of the resulting BERT model and this model is fine-tuned on the labeled source domain data to perform that aspect extraction task. This final model will eventually be applied to test data from the target domain. We next describe the two pre-training tasks of DILBERT.

Category-based Masked Language Modeling (CMLM) As noted above, Ben-David et al. (2020) integrated pivot-based training into the MLM pre-training task of BERT, in order to facilitate DA for sentiment classification. As noted in §1, it is challenging to define pivot features for the AE task, both because (a) it is a sequence-tagging, token-level task, and because (b) the aspect words and their categories tend to change across domains. Both these properties challenge the "high MI with the task label" criterion in the pivot definition (criterion (b) of § 1), as this criterion is designed for (1) sentence (or larger text) labels where words can correlate with the task label (and hence property (a) above is challenging); and where (2) the correlation with the task label in the source domain is indicative of the correlation in the target domain (and hence property (b) is challenging).

To alleviate these limitations, we present the *Category-based Masked Language Model (CMLM)* task: A variant of the BERT MLM pre-training task (left part of Figure 2). CMLM operates similarly to MLM, but with one difference: While in standard MLM the masked tokens are randomly chosen, CMLM harnesses information about the aspect categories in the source and target domains in order to mask words that are more likely to bridge the gap between the domains.[2]

We start by training static (non-contextualized) word embeddings on unlabeled data from both the source and the target domains. Then, we iterate over the input text and compute the cosine similarity between each input word and the word em-

---

[2]The aspect categories in the domains are part of the task definition, as aspect words or phrases must belong to one of these categories. For more details, see the SemEval annotation guide, section 2.1 in `https://alt.qcri.org/semeval2016/task5/data/uploads/absa2016_annotationguidelines.pdf`. Interestingly, previous work did not use this information in their algorithmic solutions.
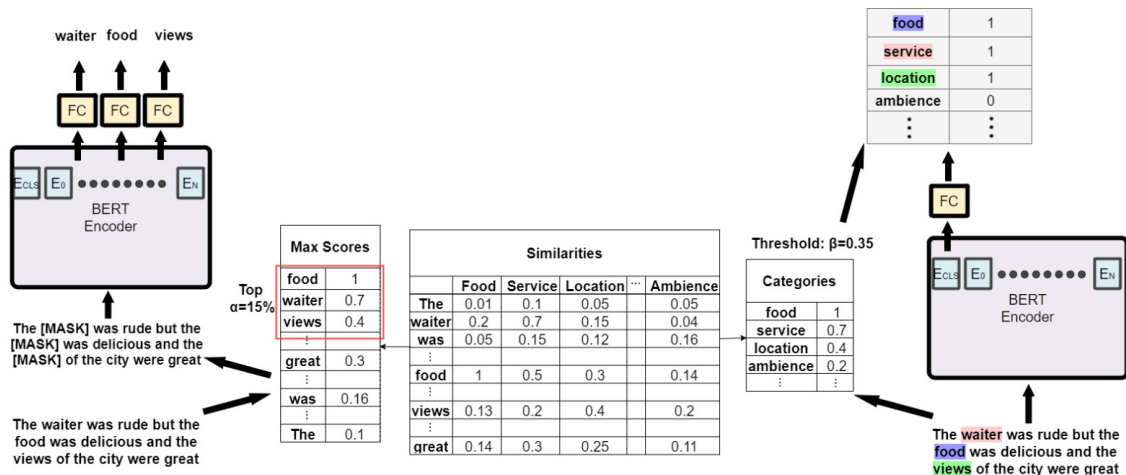
Figure 2: Illustrations of DILBERT fine-tuning tasks, CMLM and CPP, for an example input sentence. First, we calculate a similarity score for each (word, category) pair using their pre-trained, static word embeddings. Then, for each word and for each category, we keep only the maximal score. For CMLM (left), we mask the top $\alpha\%$ words, according to their score. For CPP (right), we construct a label vector in which the ith coordinate is set to 1 if the ith category score is greater than β, a threshold hyper-parameter, and to 0 otherwise. FC is a fully connected layer. The tasks are trained sequentially: First CMLM and then CPP.

bedding of each of the aspect categories from both the source and the target.[3] For each word, we keep only the highest similarity score. Once we assigned a score to each input word, we mask the top $\alpha\%$ of the input words (where α is a hyper-parameter).[4]

This masking mechanism ensures that the representation model focuses on words which can be viewed as proxy-pivots: Words that are likely to be aspect words in one of the domains. For example, given the input sentence *The pasta was delicious, I really liked it.* from the Restaurants domain, CMLM would likely mask the word *pasta*, which has a high similarity score with the *Food* aspect category, where a vanilla MLM would randomly choose a word to mask.

Category Proxy Prediction (CPP)   The CMLM task is about the prediction of masked words rather than their categories. While the masked words are strongly associated with categories, in some cases the coarser grained information that the category is represented in the text is most useful. To make our representation more category-informed, we add a second task: *Category Proxy Prediction (CPP)* (right part of Figure 2). As implied by its name, this task is about predicting the aspect categories that are represented in the input text. However, in the representation learning phase of unsupervised

DA, the unlabeled data from both domains does not have gold-standard labels of the aspect categories. We hence turn to proxy category labels instead.

More specifically, similarly to CMLM, we compute the cosine similarity between each input word and the (source and target) category names. This time, however, we keep for each category its highest score. Then we construct a binary vector, where each coordinate corresponds to one of the categories and its value is 1 if the score of that category is higher than β and 0 otherwise. Here again β is a hyper-parameter of the algorithm. Cross-entropy (which is also the loss function of CMLM) is a very natural loss for this task, as we are interested in a model that assigns high probabilities to aspect categories that are represented in the text and low probabilities to aspect categories that are not.

4   Experimental Setup

Task and Domains   We experiment with the task of cross-domain AE (Jakob and Gurevych, 2010). We consider data from the Amazon laptop reviews (L) and the Yelp restaurant reviews (R) domains. The labeled data from the L domain is taken from the SemEval 2014 task on ABSA (Pontiki et al., 2014). We follow Gong et al. (2020) and combine the SemEval 2014, 2015 and 2016 restaurant datasets (Pontiki et al., 2014, 2015, 2016) into a single dataset, removing all duplicated examples. This results in 3045/800 and 3845/2158 train/validation

---
[3]This is the static word embedding of the category name.
[4]We also considered schemes where a randomly selected sub-set of this list is masked, but observed no improvements.

examples for the L and R domains, respectively. At test time the target domain validation set serves as a test set. We obtain the list of category names for each domain from Pontiki et al. (2016).[5] We used 863,000 laptop reviews from the Amazon reviews dataset of McAuley et al. (2015), and 570,000 reviews from the Yelp open dataset [6] as our unlabeled dataset for both domains, respectively.

To consider a more challenging setup, we experiment with the MAMS dataset (Jiang et al., 2019), consisting of 5297 labeled reviews from the restaurant domain (M). We used 4297 reviews as a training set and 1000 as a validation set, and the same unlabeled data as for the R domain. Each review in the MAMS dataset has at least two aspects with different sentiment polarities, making it harder to adapt to, as the label distribution is different from that of the SemEval datasets. We consider adaptation from the L to the M domain and vice versa, adding two source-target pairs to our experiments (the M and R domains both address the same topic, and we consider them too similar for adaptation).

Experimental Protocol Focusing on unsupervised DA, we have access to unlabeled data from the source and target domains and labeled data from the source domain only. Following the protocol of representation learning for DA (§ 2), we learn a domain-invariant representation model using the unlabeled data from both domains. Then, we use the source domain labeled data to fine-tune the representation model on the downstream task. This model is eventually applied to the target domain test set data.

For DILBERT, our representation learning phase consists of two fine-tuning tasks. First, we use the large unlabeled reviews data from both domains to further train a BERT model on the CMLM task. In all our experiments (for DILBERT and the baselines) we employ the BERT-Base-Uncased model of the Hugging-Face (Wolf et al., 2019) transformers package.[7] Following Devlin et al. (2019), we fine-tune all of BERT model layers. For the CMLM task, we mask full words instead of the default sub-word masking, i.e., if we choose to mask a word, all its corresponding sub-words are masked.

Then, we use target unlabeled data to further fine-tune on the CPP task. We add a linear layer on top of the [CLS] token, and then feed forward the outputs through a sigmoid layer, using the sum of per-category binary cross-entropy losses. Once the representation learning phase is completed, we remove the CPP additional layer and the CMLM sequence classification heads. Then, we add a logistic regression head on top of all the word-level outputs and fine-tune the model on the source labeled data for the aspect. Finally, we use the fine-tuned model for predictions on the target domain test set.

To compute the similarity between words and aspect categories, we experiment with two word embedding sets: (a) The fastText pre-trained word embeddings (we refer to the DILBERT model that uses these embeddings as DILBERT-PT-WE);[8] and (b) FastText word embeddings trained on the unlabeled data from both domains (DILBERT-CT-WE; with a learning rate of 5e-5, a batch size of 4, and one training epoch.). We report results with DILBERT-CT-WE as they are better (see § 5).

For all the baselines, we keep the same protocol and design choices, unless otherwise stated.

Baselines The first baseline is UDA (Gong et al., 2020),[9] a pre-trained transformer-based neural network that performs pre-training using syntactic-driven auxiliary tasks, combined with an adversarial component (§ 2). There are two variants of UDA, differing with respect to their initialization. UDA-BASE (UDA-B) is initialized to the BERT-BASE model, while UDA-EXTENDED (UDA-E) is initialized with a BERT model which was further fine-tuned with over 20GB from the Yelp Challenge dataset and the Amazon Electronics dataset (Xu et al., 2019). As shown in Gong et al. (2020), UDA-E outperforms all previous cross-domain AE work by a large margin (§ 2.3), and hence UDA-E and UDA-B are the previous works we compare to.

To better understand the effect of our customized pre-training method, we also compare our model to a variant where everything is kept fixed except that the fine-tuning stage on the source and target unlabeled data is performed with the standard BERT model rather with DILBERT (BERT-S&T). We further compare to two similar variants that reflect a condition where we do not have access to target domain data (no domain adaptation (No-DA) setups): BERT-S, where the fine-tuning stage is performed with the standard BERT model and on unlabeled

---

[5]The category list of each domain is provided in the appendix.

[6]https://www.yelp.com/dataset

[7]https://github.com/huggingface/transformers

[8]https://github.com/facebookresearch/fastText/

[9]We use the code from the author's GitHub: https://github.com/NUSTM/BERT-UDA.

	M-L	L-M	R-L	L-R	Average
DILBERT Methods					
DILBERT	**43.72**	**58.96**	**56.07**	**61.04**	**54.95**
D-CMLM	41.97	52.74	55.48	59.36	52.39
D-CPP	31.08	29.06	36.16	40.55	34.21
Previous Baselines					
UDA-E	41.29	45.62	53.51	56.12	49.14
UDA-B	36.52	39.59	48.35	49.52	43.50
BERT-S&T	34.30	29.82	45.89	39.92	37.48
No-DA					
BERT-S	28.9	27.53	36.35	44.12	34.22
Vanilla	32.21	22.7	32.12	34.25	30.32

In Domain				
	M	L	R	Average
BERT-ID	77.87	83.8	78.82	80.16

Table 1: F1 results for both DA (top label) and in-domain (bottom table, provided for reference) setups. In the top table, column titles correspond to source-target domain pairs.

source domain data only, and Vanilla, where no unlabeled data fine-tuning is performed.

Finally, in order to understand the relative importance of the two DILBERT pre-training tasks, we experiment with the D-CMLM model (where unlabeled data fine-tuning is done with DILBERT, but only with the CMLM task), and with the D-CPP model (where unlabeled data fine-tuning is done with DILBERT, but only with the CPP task). We also evaluate the performance of an in-domain classifier (BERT-ID) – i.e., our classifier when trained on the target domain training set and evaluated on the target domain test set. This model can be seen as an upper bound on the performance we can realistically hope a DA model to achieve.

Hyper-parameter Tuning All experiments are repeated five times using different random seeds and the average results are reported. As stated, all models are based on the HuggingFace BERT-base Uncased pre-trained model. All models were tuned on the same set of hyper-parameters and the same data splits. The validation examples, used for hyper-parameter tuning, are from the source domain.[10]

5   Results and Analysis

Main Results Table 1 presents our results. As in previous work, we report the exact-match F1 score over aspect words and phrases. It shows that DILBERT, our customized pre-training procedure, outperforms all other alternatives across

[10]Full details about the hyper-parameter tuning, best hyper-parameter configurations, as well as computing infrastructure and run-time details are in Appendix A.2 and A.3.

all setups. DILBERT reduces the error of the best non-DILBERT baseline, UDA-E, by over 5% on average while using a fraction of the unlabeled data (recall that UDA-E employs a BERT model that is heavily pre-trained with data from the Restaurants and Electronic domains). The performance gap between DILBERT and UDA-E is even larger when considering the most challenging L-M setup, with over 13% improvement. The performance gaps from UDA-B, which like DILBERT is initialized with a BERT-base model, are much larger (11.45% on average across settings, 19.37% for L-M).

The comparison to the DILBERT variants that perform only one of its pre-training tasks provides a clear picture of their relative importance. Clearly, D-CMLM, the DILBERT variant which performs only the CMLM task, is an effective model, and is outperformed only by the full DILBERT. Note, however, that while the average performance gap between these models is 2.56%, in the L-M setup it is as high as 6.22%. While D-CPP is not competitive as a standalone model, its combination with D-CMLM (to form the full DILBERT model) consistently improves D-CMLM, in all DA settings. The comparison to BERT-S&T (where the representation learning stage is performed with the standard BERT rather than with DILBERT), indicates the great overall impact of DILBERT with both its tasks, as the average improvement of DILBERT is as high as 17.47% on average and 29.14% on L-M.

The performance of the No-DA baselines confirms our intuition that the MAMS restaurant domain is more challenging from a DA perspective compared to the SemEval restaurant domain (at least when adapting form/to the SemEval L domain). Additionally, and not surprisingly, BERT-S&T, which is trained on unlabeled data from both domains, outperforms BERT-S which is exposed only to source domain data.

Finally, a comparison to BERT-ID, the in-domain model which is trained and tested in the target domain, provides an indication about the performance gap that is yet to be closed. It also provides an indication of the error reduction (ER) that has already been achieved. Comparing the average performance of the Vanilla No-DA classifier to that of BERT-ID (the cross-domain error) and to that of DILBERT reveals that DILBERT cuts 24.63% from an error of 49.84% – an ER of 49.41% (for comparison, the ER of UDA-E is 37.76%).

BERT-S&T	D-CMLM
room	server
window	waiter
table	bartender
seat	waitress
person	table

(a) *There was only one MASK for the whole restaurant upstairs*. The masked word is *waiter*.

BERT-S&T	D-CMLM
that	portion
size	salad
it	bowl
this	rice
which	sandwich

(b) *MASK was just enough for me , but may not be for a big eater*. The masked word is *portions*.

MASK1		MASK2	
BERT-S&T	D-CMLM	BERT-S&T	D-CMLM
my	costumer	they	support
you	dell	i	r
this	tech	you	y
i	ace	it	ware
you	warrant	plan	us

(c) *MASK1 MASK2 would not fix the problem unless i bought your plan for $ 150 plus*. The masked words are *tech support*.

Table 2: Top 5 predictions for a given masked word according to BERT-S&T and D-CMLM, when trained on unlabeled data from the R and L domains. Examples (a) and (b) are taken from the R domain test set, and (c) from the L domain test set. For this analysis, BERT-S&T and D-CMLM were not fine-tuned for the AE task.

The Limited Unlabeled Data Scenario While the main bottleneck of adapting to a new domain is the lack of labeled data, obtaining large amounts of unlabeled data is challenging for truly resource-poor domains or languages, and may not be sufficient to learn domain-invariant representations (Ziser and Reichart, 2018a). To simulate such a scenario, we fed the DILBERT-CT-WE and DILBERT-PT-WE models (see § 4) as well as the BERT-S&T baseline with randomly sampled unlabeled data subsets from the source and the target. Figure 3 demonstrates that the two DILBERT variants outperform BERT-S&T when unlabeled data is scarce, and in some cases, DILBERT outperforms BERT-S&T even when using less unlabeled data (e.g., DILBERT with 35MB of unlabeled data from each domain, compared to BERT-S&T with 70MB).
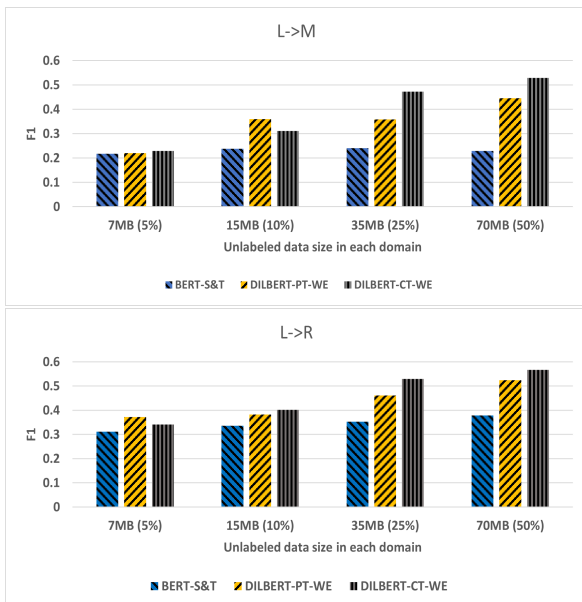

Figure 3: AE performance as a function of the unlabeled dataset size used for representation learning.

CMLM Probing We would next like to shed more light on the quality of the CMLM task, which has the strongest impact on DILBERT's results. For this aim, we fine-tune the D-CMLM and BERT-S&T models on the unlabeled data of the R and L domains, and apply them to test sentences from these domains, without task-related (AE) fine-tuning with label data. Table 2 presents the words predicted by these models for three representative masking tasks. Obviously, the predictions of D-CMLM are much more semantically related to the masked tokens. While this is a qualitative analysis, limited in nature to a small number of examples, our manual inspection of the results suggests that this pattern is the rule rather than the exception.

Generalizing Beyond Category Names We would further like to verify that by considering the information encoded in the category names DILBERT can better identify aspect words and phrases that are not identical to one of these category names. We hence re-evaluate all models such that aspect words that are also category names (e.g., in the restaurant review sentence: *The food was great* where *food* is both an aspect word and the name of its category) are not considered in the evaluation.[11]

Table 3 reports the results of this evaluation. The observed patterns are similar to the main evaluation results: DILBERT is still the best DA model by a large margin. Not surprisingly, the absolute results of all models are lower than in the main evaluation as the excluded aspect words are more typical and hence easier for the models to identify.

Standard Deviations Finally, Table 4 reports the standard deviations for all models across the five

---

[11]For more details about this scenario – the category names and their prevalence among the aspect terms in the texts, see appendix A.1.

	M-L	L-M	R-L	L-R	Average
DILBERT Methods					
DILBERT	**36.18**	**48.35**	**48.69**	**43.81**	**44.26**
D-CMLM	35.34	46.21	48.18	43.32	43.26
D-CPP	23.06	16.94	31.22	27.72	24.74
Previous Baselines					
UDA-E	32.74	37.37	46.13	39.54	38.95
UDA-B	30.73	29.71	38.06	32.38	32.72
BERT-S&T	25.87	22.78	35.77	31.35	28.94
No-DA					
BERT-S	21.43	18.14	28.85	29.03	24.36
Vanilla	25.45	18.76	29.14	24.79	24.53

In Domain				
	M	**L**	**R**	**Average**
BERT-ID	69.7	75.19	67.48	70.79

Table 3: The results of a re-evaluation of all the models where aspect words that are identical to category names are excluded from the evaluation.

	M-L	L-M	R-L	L-R	Average
DILBERT Methods					
DILBERT	1	**0.4**	1.3	1.8	**1.1**
D-CMLM	2	0.95	**1.1**	1	1.26
D-CPP	**0.62**	5.03	2.84	3.95	3.11
Previous Baselines					
UDA-E	1.03	1.85	0.6	2.17	1.41
UDA-B	1.51	1.68	1.54	2.29	1.76
BERT-S&T	2.08	4.3	2.17	1.39	1.85
No-DA					
BERT-S	1.45	2.39	2.17	1.39	1.85
Vanilla	1.23	2.63	2.2	1.47	1.88

In Domain				
	M	**L**	**R**	**Average**
BERT-ID	0.33	0.77	0.41	0.51

Table 4: Standard deviations for all models across the five folds of the cross-validation protocol.

folds of the cross-validation protocol. DILBERT has the lowest averaged standard-deviation among all DA models and No-DA baselines. This is obviously another advantage of our proposed model, particularly that in unsupervised domain adaptation there is no labeled target domain data available for model selection, and hence stability to random seeds is crucial (Ziser and Reichart, 2019).

6 Conclusions and Future Work

We have presented DILBERT, a customized pre-training approach for unsupervised DA with category shift, and apply it to the task of aspect extraction. We demonstrate that by fine-tuning with a modified version of the BERT pre-training tasks, we can better adapt to new domains and aspect categories, even in resource-poor scenarios where unlabeled data is limited. To make our experimentation more challenging, we presented a new AE domain (M), which is substantially different from previously presented ones.

In future work, we would like to extend our approach so that it can jointly solve the aspect extraction and sentiment analysis tasks (the ABSA task). Moreover, we would like to verify the quality of our approach in additional domains, tasks (i.e., going beyond AE to other tasks that present category shifts when domains change) and eventually even languages.

Acknowledgments

References

Eyal Ben-David, Carmel Rabinovitz, and Roi Reichart. 2020. PERL: Pivot-based domain adaptation for pre-trained deep contextualized embedding models. *Transactions of the Association for Computational Linguistics*, 8:504–521.

Shai Ben-David, John Blitzer, Koby Crammer, Fernando Pereira, et al. 2007. Analysis of representations for domain adaptation. In *Advances in neural information processing systems*.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447, Prague, Czech Republic. Association for Computational Linguistics.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 120–128.

Minmin Chen, Kilian Q Weinberger, and Yixin Chen. 2011. Automatic feature decomposition for single view co-training. In *ICML*.

Minmin Chen, Zhixiang Eddie Xu, Kilian Q Weinberger, and Fei Sha. 2012. Marginalized denoising autoencoders for domain adaptation. In *ICML*.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic. Association for Computational Linguistics.

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, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ying Ding, Jianfei Yu, and Jing Jiang. 2017. Recurrent neural networks with auxiliary labels for cross-domain opinion target extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *ICLR*.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030.

Chenggong Gong, Jianfei Yu, and Rui Xia. 2020. Unified feature and instance based domain adaptation for aspect-based sentiment analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7035–7045, Online. Association for Computational Linguistics.

Niklas Jakob and Iryna Gurevych. 2010. Extracting opinion targets in a single and cross-domain setting with conditional random fields. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 1035–1045.

Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. 2019. A challenge dataset and effective models for aspect-based sentiment analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6280–6285, Hong Kong, China. Association for Computational Linguistics.

Wei Jin, Hung Hay Ho, and Rohini K Srihari. 2009. A novel lexicalized hmm-based learning framework for web opinion mining. In *Proceedings of the 26th annual international conference on machine learning*.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430, Sapporo, Japan. Association for Computational Linguistics.

Fangtao Li, Chao Han, Minlie Huang, Xiaoyan Zhu, Ying-Ju Xia, Shu Zhang, and Hao Yu. 2010. Structure-aware review mining and summarization. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 653–661, Beijing, China. Coling 2010 Organizing Committee.

Zheng Li, Xin Li, Ying Wei, Lidong Bing, Yu Zhang, and Qiang Yang. 2019. Transferable end-to-end aspect-based sentiment analysis with selective adversarial learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4590–4600, Hong Kong, China. Association for Computational Linguistics.

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*, pages 1433–1443, Lisbon, Portugal. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard S Zemel. 2016. The variational fair autoencoder. In *ICLR*.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. 2008. Domain adaptation with multiple sources. *Advances in neural information processing systems*, 21:1041–1048.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 43–52.

Timothy Miller. 2019. Simplified neural unsupervised domain adaptation. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2019, page 414. NIH Public Access.

Oren Pereg, Daniel Korat, and Moshe Wasserblat. 2020. Syntactically aware cross-domain aspect and opinion terms extraction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1772–1777.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit.

2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, Denver, Colorado. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. 2011. Opinion word expansion and target extraction through double propagation. *Computational Linguistics*, 37(1):9–27.

Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in nlp—a survey. *arXiv preprint arXiv:2006.00632*.

Tun Thura Thet, Jin-Cheon Na, and Christopher SG Khoo. 2010. Aspect-based sentiment analysis of movie reviews on discussion boards. *Journal of information science*, 36(6):823–848.

Hao Tian, Can Gao, Xinyan Xiao, Hao Liu, Bolei He, Hua Wu, Haifeng Wang, and Feng Wu. 2020. SKEP: Sentiment knowledge enhanced pre-training for sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4067–4076, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Wenya Wang and Sinno Jialin Pan. 2018. Recursive neural structural correspondence network for cross-domain aspect and opinion co-extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2171–2181, Melbourne, Australia. Association for Computational Linguistics.

Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2016. Recursive neural conditional random fields for aspect-based sentiment analysis. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 616–626, Austin, Texas. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2019. Bert post-training for review reading comprehension and aspect-based sentiment analysis.

Yftah Ziser and Roi Reichart. 2017. Neural structural correspondence learning for domain adaptation. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 400–410, Vancouver, Canada. Association for Computational Linguistics.

Yftah Ziser and Roi Reichart. 2018a. Deep pivot-based modeling for cross-language cross-domain transfer with minimal guidance. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 238–249, Brussels, Belgium. Association for Computational Linguistics.

Yftah Ziser and Roi Reichart. 2018b. Pivot based language modeling for improved neural domain adaptation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1241–1251, New Orleans, Louisiana. Association for Computational Linguistics.

Yftah Ziser and Roi Reichart. 2019. Task refinement learning for improved accuracy and stability of unsupervised domain adaptation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5895–5906, Florence, Italy. Association for Computational Linguistics.

A Appendix

A.1 Aspect Categories

Table 5 provides the category names of our three domains, which were available to our model. There are 28 categories for the laptops (L) domain and 9 categories for the restaurants domains (R and M). In the R test set there are 876 unique aspect terms (i.e., unique words or phrases that are annotated as aspects), and 21.1% of these are words that are identical to one of the category names. In the M test set the corresponding numbers are 835 unique aspect terms, and 12.4% of these are words that are identical to one of the category names. In the laptops (L) test set the corresponding numbers are 387 unique aspect terms, and 12.7% of these are words that are identical to one of the category names.

Laptops	Restaurants
cooling, multimedia,warranty, OS, portability, motherboard, support, keyboard, quality, powersupply, ports, fan, software, memory, battery, graphics, display, usability, price, design, connectivity, performance, shipping, laptop, disk, hdd, cpu, mouse	drinks, food, service, prices, style, location, ambience, quality, restaurant

Table 5: The list of aspect categories of each of the participating domains. *Restaurants* refer both to the R and the M domains.

A.2 Computing Infrastructure, Run-time and Number of Parameters

All models were trained on a Quadro RTX 6000 machine, using 1 GPU card. The average run-time for the CMLM training was about 2 hours. The average run-time for the CPP training was about 6 minutes. The average run-time for the AE training was about 6 minutes. The inference time on the test set for the AE task was about 1 minute. The BERT-base model has 109.5M parameters. The CMLM architecture has additional 620K parameters. The CPP architecture has additional 21.5K parameters. The AE task classifier has additional 2307 parameters.

A.3 Hyper-parameter Tuning

All models are based on the HuggingFace BERT-base Uncased pre-trained model. We use their default word-piece vocabulary, and the AdamW optimizer (Loshchilov and Hutter, 2018) with an $\epsilon = 1e - 8$ and a linearly decreasing learning rate schedule.

The number of words to mask (α) in the CMLM task was chosen among 5%, 10%, and 15% of the review length. For the CPP task, the number of epochs was chosen among {1, 2, 3}, and the β threshold among $[0.25, 0.251, \ldots, 0.45]$. The learning rate was 5e-5 and the batch size was 8.

In the AE classification stage, the number of epochs was chosen among {2,3,4}, the batch size among {16, 32, 64} and the learning rate among {2e-5, 3e-5, 5e-5}.

For both UDA-B and UDA-E, the hyper-parameter search of Gong et al. (2020) is performed over a sub-set of the set we consider here. We hence re-run all of their experiments with our grid search, which led to better performance than reported in their work.

Best Hyper-parameters The hyper-parameters of the best DILBERT configuration across cross-validation folds and DA setups were as follows:

- CMLM, α: 10%.

- CPP, number of epochs : 1.

- CPP, β:
  R-L: 0.353, L-R, L-M: 0.406, M-L: 0.362.

- AE (task classifier), number of epochs: 3

- AE (task classifier), batch size: 32.

- AE (task classifier), learning rate: 5e-5.