Cross-Domain Review Generation for Aspect-Based Sentiment Analysis

Jianfei Yu, Chenggong Gong, and Rui Xia[*]
School of Computer Science and Engineering,
Nanjing University of Science and Technology, China
{jfyu, cggong, rxia}@njust.edu.cn

Abstract

Supervised learning methods have proven to be effective for Aspect-Based Sentiment Analysis (ABSA). However, the lack of fine-grained labeled data hinders their effectiveness in many domains. To address this issue, unsupervised domain adaptation methods are desired to transfer knowledge from a labeled source domain to any unlabeled target domain. In this paper, we propose a new domain adaptation paradigm called cross-domain review generation (CDRG), which aims to generate target-domain reviews with fine-grained annotation based on the source-domain labeled reviews. To achieve this goal, we propose a two-step approach as a concrete realization of CDRG. It first converts a source-domain review to a domain-independent review by masking its source-specific attributes, and then converts the domain-independent review to a target-domain review with a masked language model pre-trained in the target domain. We further propose two ways to leverage the generated target-domain reviews for two cross-domain ABSA tasks. Extensive experiments demonstrate the superiority of our CDRG-based approaches over the state-of-the-art domain adaptation methods.

1 Introduction

Aspect-Based Sentiment Analysis (ABSA) has received considerable attention in recent years (Liu, 2020). The goal of ABSA is to extract the aspect terms mentioned in review sentences and predict the sentiments over them (Pontiki et al., 2016). For example, given the review "*The fish soup is delicious*", the aspect term and its sentiment are *fish soup* and *Positive*, respectively. With the recent advances in deep learning, many supervised neural models have been proposed for several ABSA tasks, e.g., aspect extraction (Liu et al., 2015; Xu et al.,

2018), aspect-level sentiment classification (Wang et al., 2016b; Tang et al., 2016), and End-to-End ABSA (Zhang et al., 2015; Li et al., 2019a).

Although these neural models have obtained promising results in several product domains such as *Laptop* and *Restaurant* (Pontiki et al., 2014), the major obstacle to them is the lack of rich annotated resources in many new domains. Since ABSA requires fine-grained annotation of aspect terms and their sentiments in each review, it is extremely time-consuming to develop such resources for each domain, and the annotation process can be prohibitively expensive. To alleviate the annotation efforts, unsupervised domain adaptation methods are desired to transfer knowledge from a source domain with rich labeled data to a target domain with only unlabeled data.

The key challenge of domain adaptation is that the data distribution of the source domain usually differs from that of the target domain. To alleviate the domain discrepancy, many approaches have been proposed in coarse-grained sentiment classification based on the following two paradigms:

- **feature-based adaptation**, which aims to aims to learn a domain-invariant feature representation across domains (Blitzer et al., 2007; Glorot et al., 2011; Yu and Jiang, 2017; Ziser and Reichart, 2018; Ghosal et al., 2020)
- **instance-based adaptation**, which focuses on re-weighting labeled instances in the source domain for use in the target domain (Mansour et al., 2008; Dredze et al., 2010).

However, due to the challenges in fine-grained adaptation, only a few studies explore the domain adaptation problem for ABSA (Ding et al., 2017; Wang and Pan, 2018; Li et al., 2019b; Gong et al., 2020). Moreover, these studies still follow the above two domain adaptation paradigms, which suffer from two common limitations: (1) although
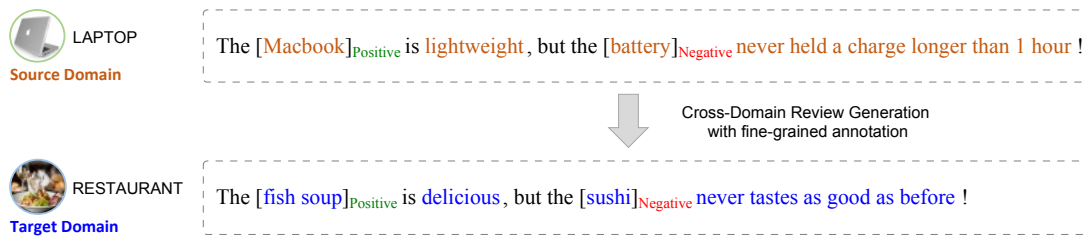
---

[*]Corresponding author.

Figure 1: Cross-Domain Review Generation with applications to the End-to-End ABSA task, in which the goal is to generate a target-domain review with fine-grained annotation given a labeled review in the source domain.

they can reduce the domain discrepancy by learning shared representations or re-weighting source instances, the supervision signals for their main task solely come from the labeled source domain; (2) both of them are lack of interpretability, as the shared representations or re-weighted instances offer little transparency regarding the knowledge transferred to the target domain.

To address the two limitations, we propose a new domain adaptation paradigm named **Cross-Domain Review Generation** (CDRG) with applications to the ABSA task. Given a labeled review in the source domain, the goal is to generate a target-domain review with fine-grained annotation, which converts the domain-specific attributes (e.g., aspects, opinions, and collocations) to the target domain while preserving its annotation and remaining contents. For example, in Fig. 1, a labeled review from the *Laptop* domain is transferred to the *Restaurant* domain by converting its source-specific attributes to target-specific attributes (e.g., *Macbook* to *fish soup*, *lightweight* to *delicious*, etc) but keeping other words and the labels unchanged. Different from existing text generation tasks, CDRG is challenging due to a couple of reasons. First, there is no paralleled corpus which aligns labeled source-domain reviews and unlabeled target-domain reviews. Second, given unparalleled corpus, it is non-trivial to achieve alignments between the domain-specific attributes. More importantly, the generated target review is required to have the fine-grained annotation.

To achieve this goal, we propose a simple yet effective two-step approach, containing a domain generalization step and a domain specification step. Specifically, the domain generalization step first identifies important domain-specific attributes[1],

and then mask source-specific attributes in each source review to obtain a domain-independent review. Next, given the domain-independent review as input, the domain specification step employs a pre-trained masked language model from the target domain to generate a target review. In our two-step approach, the domain-independent review serves as a bridge to achieve word-to-word alignments in source and target reviews, and thus the fine-grained annotation from the source review can be directly transferred to the target review.

We further propose two training strategies to leverage the generated target-domain reviews for two cross-domain ABSA tasks, including cross-domain End-to-End ABSA (E2E-ABSA) and aspect extraction (AE). Experiment results on four benchmark datasets show that only using our generated target-domain reviews, the baseline BERT model already outperforms the state-of-the-art domain adaptation methods, and a joint usage of generated target reviews and labeled source reviews can further boost the performance significantly.

Our main contributions can be summarized as follows:

- We propose a new domain adaptation paradigm named **Cross-Domain Review Generation** (CDRG) with applications to the ABSA task, and then devise a simple yet effective two-step approach as a concrete realization of CDRG.

- With the help of generated target-domain reviews, our best training strategy outperforms the state-of-the-art method by an absolute improvement of 2.83% and 4.47% on Micro-F1 for cross-domain E2E-ABSA and cross-domain AE, respectively.

- As long as a source domain has sufficient annotated reviews, our two-step approach for CDRG can generate many annotated target reviews, which offer interpretable justification for domain adaptation.

---

[1]Domain-specific attributes can be words, phrases, syntactic structures, and expression styles only occurring in the source domain or the target domain. However, as aspect and opinion terms are the core of ABSA, we only consider domain-specific aspect and opinion terms in our approach.

2 Related Work

As two important tasks in ABSA, aspect extraction (Liu et al., 2015; Poria et al., 2016; Wang et al., 2016a, 2017; Li et al., 2018a; Xu et al., 2018) and aspect-level sentiment classification (Dong et al., 2014; Tang et al., 2016; Wang et al., 2016b; Yang et al., 2017; Ma et al., 2017; Wang et al., 2018) have been extensively studied in the literature. For practical applications, a number of recent studies handle them together in an end-to-end manner, in which many supervised learning methods with discrete linear features (Mitchell et al., 2013) and continuous neural features (Zhang et al., 2015; Li et al., 2019a) have been proposed. Despite obtaining promising results, their main limitation lies in the lack of annotated data in many new domains. To address this data sparsity problem, unsupervised domain adaptation methods are desired.

Most existing domain adaptation studies focus on coarse-grained sentiment classification to learn domain-invariant representations, including pivot-based methods (Blitzer et al., 2007; Pan et al., 2010; Yu and Jiang, 2016), auto-encoders (Chen et al., 2012; Zhuang et al., 2015), semi-supervised methods (He et al., 2018; Ye et al., 2020), and domain adversarial networks (Ganin et al., 2016; Li et al., 2018b). Besides, another line of work focuses on re-weighting source instances to automatically find useful source samples for the target domain (Jiang and Zhai, 2007; Mansour et al., 2008; Dredze et al., 2010; Xia et al., 2014). Due to the challenges in fine-grained adaptation, there exist only a few studies for cross-domain aspect and opinion extraction (Li et al., 2012; Ding et al., 2017; Wang and Pan, 2018), or End-to-End ABSA (Li et al., 2019b; Gong et al., 2020). However, these methods still follow the traditional domain adaptation paradigms to either learn shared representations or perform instance weighting. Different from these methods, we propose to accomplish domain adaptation for ABSA based on Cross-Domain Review Generation.

3 Problem Formulation

In this paper, we consider two ABSA tasks, i.e., End-to-End ABSA (E2E-ABSA) and Aspect Extraction (AE). Following Li et al. (2019b), we formulate both tasks as sequence labeling problems. Formally, given a sequence of input tokens $\mathbf{x} = \{w_1, w_2, ..., w_n\}$, its label sequence is denoted by $\mathbf{y} = \{y_1, y_2, ..., y_n\}$. Let $y_j \in \{\texttt{B-POS}, \texttt{I-POS},$ $\texttt{B-NEG}, \texttt{I-NEG}, \texttt{B-NEU}, \texttt{I-NEU}, \texttt{O}\}$ be the label space for the E2E-ABSA task, and $y_j \in \{\texttt{B}, \texttt{I}, \texttt{O}\}$ be the label space for the AE task.

Cross-Domain ABSA. We focus on the unsupervised domain adaptation setting, in which labeled data are only available from the source domain. Specifically, we assume access to a set of labeled reviews from the source domain $D_S = \{(\mathbf{x}_i^s, \mathbf{y}_i^s)\}_{i=1}^{N^s}$, and another set of unlabeled reviews from the target domain $D_U = \{\mathbf{x}_i^u\}_{i=1}^{N^u}$. The goal is to predict the label sequence for test data in the target domain: $\mathbf{y}_i^t = f_t(\mathbf{x}_i^t)$, $D_T = \{\mathbf{x}_i^t\}_{i=1}^{N^t}$.

4 Methodology

Overview. In this work, we propose a new domain adaptation paradigm named Cross-Domain Review Generation (CDRG). The goal is to transform a labeled source review to a labeled target review by converting its source-specific attributes to target-specific attributes while retaining the remaining contents and its labels. To achieve this goal, we propose a simple yet effective two-step approach, as shown in Fig. 2. In the domain generalization step, we first extract the domain-specific attributes from the labeled source data and the unlabeled target data. For each labeled source review, we mask its source-specific attributes to obtain a domain-independent review. In the domain specification step, a pre-trained BERT model is first re-trained with masked language modeling (MLM) on the unlabeled data from the target domain. Next, given the domain-independent review, we employ the target-domain MLM to generate a target review. As the domain-independent review serves as a bridge to achieve word-to-word alignments between the source and generated target reviews, the source labels can be directly transferred to the target review. Finally, we propose two strategies to leverage labeled target reviews for E2E-ABSA and AE tasks, including independent training and merge training.

4.1 Proposed Two-Step Approach for CDRG

Our proposed approach for CDRG consists of two steps: (1) a domain generalization step to convert each source-domain review to a domain-independent review; (2) a domain specification step to convert a domain-independent review to a target-domain review.
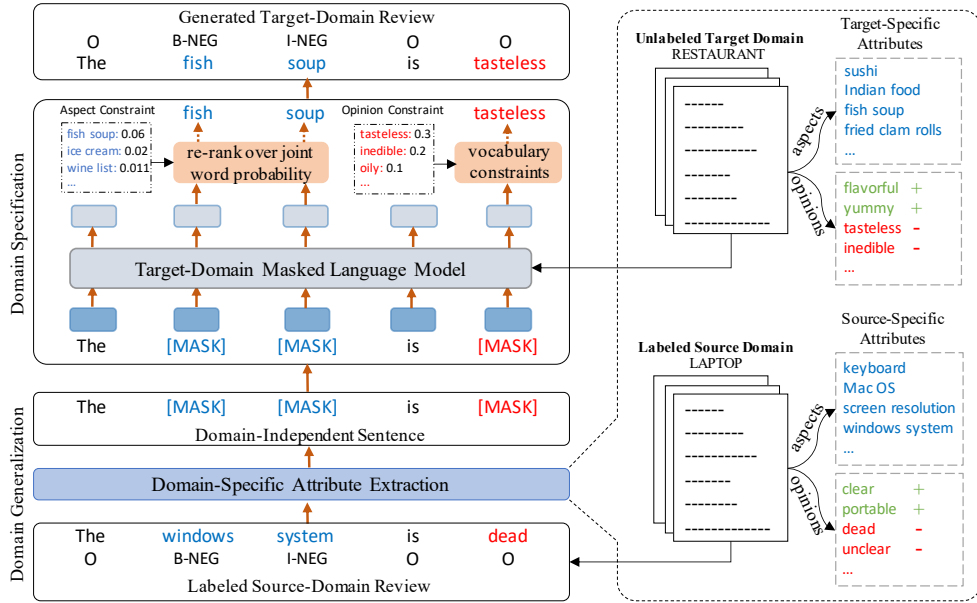
Figure 2: Overview of our two-step approach for Cross-Domain Review Generation, which contains a domain generalization step and a domain specification step. Note that we use Double Propagation to extract all the opinion terms together with their sentiments, and +, -, and 0 denote Positive, Negative, and Neutral sentiments, respectively.

4.1.1 Step 1: Domain Generalization

To convert each source-domain review to a domain-independent review, we first extract the domain-specific attributes from each domain, and then mask the source-specific attributes in source reviews.

Domain-Specific Attribute Extraction. We define domain-specific attributes as words, phrases, syntactic structures, and expression styles that only occur in the source domain or the target domain. However, it is often challenging to identify some attributes such as syntactic structures and expression styles, as most of them are implicitly expressed in the review. More importantly, since ABSA aims to jointly extract the aspects and sentiments, aspect and opinion terms tend to play more crucial roles than the other attributes. Therefore, we only consider domain-specific aspect terms and opinion terms as domain-specific attributes in this paper.

To extract aspect and opinion terms from the unlabeled target domain, we use a dependency relation-based unsupervised method named Double Propagation (Qiu et al., 2011). Specifically, given the unlabeled target reviews D_U, we first resort to a sentiment lexicon[2] to extract the opinion terms in D_U, and use the *conj* relation to expand the opinion term list. Next, we employ these opinion terms as seed words, and extract all the words holding the *amod* and *nsubj* relations towards any seed word.

We then treat the extracted words as aspect terms[3], and use the *nn* relation to expand the aspect term list. The above three steps can be iterated until the aspect and opinion term lists are no longer updated.

Given the labeled source reviews D_S, we can easily obtain the aspect term list, as the aspect terms have been annotated in each review for both E2E-ABSA and AE tasks. For opinion terms, we also utilize Double Propagation to expand its list.

After obtaining the aspect and opinion term lists for each domain, we remove all the domain-independent terms that occur in both source and target domains, and obtain the domain-specific term lists. Let us use A_s and O_s to denote the source-specific aspect and opinion term lists, and A_t and O_t the target-specific aspect and opinion term lists.

Domain-Independent Reviews. Given each review in the source domain $\mathbf{x}^s = \{w_1, w_2, ..., w_n\}$, if its sub-sequence is a source-specific attribute in either A_s or O_s, we substitute each word in the sub-sequence with a special token [MASK] to obtain a domain-independent review, denoted by $\mathbf{x}^m$. For example, in Fig 2, the source-specific aspect term *windows system* and opinion term *dead* are replaced by two [MASK] tokens and one [MASK] token, respectively.

---

[2] https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon

[3] Since Double Propagation can only extract single-word aspect terms, we further expand them to multi-word aspect terms if their previous words have an *amod* or *nn* relation towards them (e.g., *Italian* $\xrightarrow{amod}$ *food*, *fish* $\xrightarrow{nn}$ *soup*, etc.).

4.1.2 Step 2: Domain Specification

The domain specification step is responsible for incorporating target-specific attributes into each domain-independent review $\mathbf{x}^m$. In our work, we propose to transform it as a text infilling problem. To address this, we pre-train a Transformer network with the masked language modeling (MLM; Devlin et al. (2019)) task on unlabeled data in the target domain, followed by predicting a target-specific word for each masked token in $\mathbf{x}^m$.

Target-Domain MLM. Since the size of the unlabeled target-domain data D_U is typically small, we adopt a pre-trained BERT model (Devlin et al., 2019) and re-train the BERT model with MLM on D_U. Specifically, we create training instances by randomly replacing a subset of tokens with [MASK] in each unlabeled review $\mathbf{x}^u$, and the objective is to recover the masked tokens based on the hidden states from BERT.

With the target-domain MLM (TD-MLM), we can infill each masked position in $\mathbf{x}^m$ based on their context. Let us use $M = \{m_1, m_2, ..., m_K\}$ to denote the indexes of the masked tokens in $\mathbf{x}^m$, where K refers to the number of masked tokens. The predicted word for the j-th masked token can be computed as follows:

$$p(w_{m_j}|\mathbf{x}^m) = \text{TD-MLM}(w_{m_j}|\mathbf{x}^m);$$
$$o_{m_j} = \arg\max_{w_{m_j}} p(w_{m_j}|\mathbf{x}^m), \quad w_{m_j} \in V, \quad (1)$$

where V denotes the whole vocabulary, and o_{m_j} is the output word with the highest probability.

However, the TD-MLM suffers from two major limitations: (1) the predicted token with the highest probability in Eqn. (1) may not be an aspect term or an opinion term; (2) each masked token is predicted independently, and thus it is possible that the predicted tokens for two consecutive masked positions are not coherent.

Target-Specific Aspect Constraint. To tackle these two limitations, we first propose to utilize the target-specific aspect terms (i.e., A_t) extracted in Section 4.1.1 as vocabulary constraints to limit the prediction space of each masked aspect term.

Specifically, if the masked aspect term corresponds to a single-word term in the source review (e.g., *keyboard*), the word selection in Eqn. (1) can be modified as follows:

$$o_{m_j} = \arg\max_{w_{m_j}} p(w_{m_j}|\mathbf{x}^m), \quad w_{m_j} \in A_t^1, \quad (2)$$

where A_t^1 refers to the sets of single-word aspect terms in A_t. Otherwise, if the masked aspect term

corresponds to a multi-word term, we compute the joint word probabilities of each multi-word term in A_t followed by re-ranking them. Let us use k to denote the number of consecutive masked tokens, and $m_{j:j+k}$ to denote the span of the masked aspect term. The word selection for k consecutive masked tokens can thus be computed as follows:

$$p(w_{m_{j:j+k}}|\mathbf{x}^m) = \prod_{i=j}^{j+k-1} p(w_{m_i}|\mathbf{x}^m);$$
$$o_{m_{j:j+k}} = \arg\max_{w_{m_{j:j+k}}} p(w_{m_{j:j+k}}|\mathbf{x}^m), \quad (3)$$

where $w_{m_{j:j+k}} \in A_t^k$, and A_t^k refers to the sets of k-word aspect terms in A_t.

Target-Specific Opinion Constraint. For each masked opinion term, it is important to keep the sentiment consistency when predicting its corresponding target-specific opinion term.

To achieve this, we resort to the Double Propagation algorithm, which relies on the sentiment lexicon to assign sentiment (i.e., Positive, Negative, and Neutral) to each aspect term and opinion term in a bootstrapping manner. Based on the output, we obtain the sentiment of all the source-specific and target-specific opinion terms (i.e., O_s and O_t) extracted in Section 4.1.1.

Next, we look up the sentiment of the masked source-specific opinion term, and then utilize all the target-specific opinion terms with the same sentiment in O_t as vocabulary constraints in Eqn. (2) and Eqn. (3) to generate the target-specific opinion terms. For example, in Fig 2, since *dead* is a source-specific negative opinion term, we use all the single-word target-specific opinion terms with the negative sentiment as vocabulary constraints in Eqn. (2) to generate the opinion term *tasteless*.

Generated Target-Domain Reviews. Based on Eqn. (2) and Eqn. (3), we can infill the masked positions in each domain-independent review $\mathbf{x}^m$, and obtain the generated target-domain review, denoted by $\mathbf{x}^g$. It is worth noting that if a source-domain review $\mathbf{x}^s$ does not contain any source-specific attributes, its generated target-domain review $\mathbf{x}^g$ will be the same as $\mathbf{x}^s$.

With the domain-independent review $\mathbf{x}^m$, our two-step approach essentially achieves word-to-word alignments between $\mathbf{x}^s$ and $\mathbf{x}^g$. Therefore, we can directly employ the sequence label of $\mathbf{x}^s$ as the fine-grained annotation of $\mathbf{x}^g$. Fig. 2 shows a label transferring example for the E2E-ABSA task.

Formally, we use $D_G = \{(\mathbf{x}_i^g, \mathbf{y}_i^s)\}_{i=1}^{N^s}$ to denote the set of the generated target-domain reviews.

Dataset	Domain	Sentences	Training	Testing
L	Laptop	3845	3045	800
R	Restaurant	6035	3877	2158
D	Device	3836	2557	1279
S	Service	2239	1492	747

Table 1: Basic statistics of our datasets.

4.2 Post-Generation Training for Main Tasks

After obtaining the set of generated target-domain reviews D_G, we further propose two strategies to leverage them to train effective E2E-ABSA and AE models for the target domain as follows.

Independent Training. An intuitive strategy is to solely treat D_G as training instances, and directly train a sequence labeling model over them. Following Gong et al. (2020), we adopt the pre-trained BERT model as the text encoder, followed by fine-tuning it on D_G.

Merge Training. Since the qualities of the generated target-domain reviews rely on the target-domain MLM as well as the aspect and opinion terms extracted by Double Propagation, it is inevitable that D_G contains a number of aspect and opinion terms with incorrect annotations. Therefore, we propose to merge the labeled source reviews D_S with D_G as the training instances, which may alleviate the annotation noise in D_G. Similar to Independent Training, a BERT-based sequence labeling model is trained over the merged corpus $D_S \cup D_G$.

5 Experiment

5.1 Experiment Settings

Datasets. We use four benchmark datasets including Laptop (L), Restaurant (R), Device (D), and Service (S) for experiments. L is from SemEval-2014 ABSA challenge (Pontiki et al., 2014), containing user reviews from the laptop domain. R refers to a combination of the restaurant datasets from SemEval ABSA challenge 2014, 2015, and 2016 (Pontiki et al., 2014, 2015, 2016). D is the union set of all the digital device reviews collected by Toprak et al. (2010). S contains reviews from web services, which is introduced by Hu and Liu (2004). The basic statistics are shown in Table 1.

Settings. We carry out experiments on 10 transfer pairs with the four domains above. Following previous work (Li et al., 2019b; Gong et al., 2020), we remove D→L and L→D, as the two domains are very similar. For each transfer pair, the training

data is a combination of the labeled training data in the source domain and the unlabeled training data in the target domain. We report the evaluation results on the test data from the target domain. For fair comparison with previous work, we use the Micro-F1 score with the exact match as the evaluation metric, which means that the predicted labels can be counted as correct only if they are exactly matched with the golden labels.

5.2 Compared Systems & Hyperparameters

To show the effectiveness of our Cross-Domain Review Generation (CDRG)-based methods, we consider the following compared systems:

- DP (Qiu et al., 2011): the unsupervised Double Propagation method detailed in Section 4.1.1.

- Hier-Joint (Ding et al., 2017): An LSTM-based domain adaptation method with syntactic rule-based auxiliary tasks for cross-domain AE.

- RNSCN (Wang and Pan, 2018): A recursive neural structural correspondence network based on syntactic structures and auto-encoders.

- AD-SAL (Li et al., 2019b): A Selective Adversarial Learning method to achieve local semantic alignments for fine-grained domain adaptation.

- $BERT_B$ and $BERT_E$: directly fine-tuning two versions of pre-trained encoders on the labeled source domain. $BERT_B$ is the uncased $BERT_{base}$ model from Devlin et al. (2019), and $BERT_E$ is another uncased $BERT_{base}$ model from Xu et al. (2019), which is pre-trained on E-commerce reviews from the Amazon Electronics dataset (He and McAuley, 2016) and the Yelp Challenge.

- $BERT_B$-UDA and $BERT_E$-UDA (Gong et al., 2020): our recent unified feature and instance-based domain adaptation method based on $BERT_B$ and $BERT_E$, respectively.

Besides, we consider the following variants of our CDRG-based domain adaptation methods:

- $BERT_B$-CDRG-X: re-training $BERT_B$ with the MLM task to obtain the Target-Domain MLM in Section 4.1.2, and employing $BERT_B$ as the base model for the two training strategies in Section 4.2. Here X can be independent training and merge training.

- $BERT_E$-CDRG-X: replacing the $BERT_B$ model in $BERT_B$-CDRG-X with $BERT_E$.

Methods	S→R	L→R	D→R	R→S	L→S	D→S	R→L	S→L	R→D	S→D	Average
DP (Qiu et al., 2011)	34.47	34.47	34.47	18.31	18.31	18.31	16.63	16.63	19.03	19.03	22.97
Hier-Joint (Ding et al., 2017)	31.10	33.54	32.87	15.56	13.90	19.04	20.72	22.65	24.53	23.24	23.72
RNSCN (Wang and Pan, 2018)	33.21	35.65	34.60	20.04	16.59	20.03	26.63	18.87	33.26	22.00	26.09
AD-SAL (Li et al., 2019b)	41.03	43.04	41.01	28.01	27.20	26.62	34.13	27.04	35.44	33.56	33.71
BERT$_B$	44.66	40.38	40.32	19.48	25.78	30.31	31.44	30.47	27.55	33.96	32.44
BERT$_B$-UDA (Gong et al., 2020)	47.09	45.46	42.68	33.12	27.89	28.03	33.68	34.77	34.93	32.10	35.98
BERT$_B$-CDRG-Indep Training	44.46	44.96	39.42	34.10	33.97	31.08	33.59	26.81	25.25	29.06	34.27
BERT$_B$-CDRG-Merge Training	47.92	49.79	47.64	35.14	38.14	37.22	38.68	33.69	27.46	34.08	38.98
BERT$_E$	51.34	45.40	42.62	24.44	23.28	28.18	39.72	35.04	33.22	33.22	35.65
BERT$_E$-UDA (Gong et al., 2020)	**53.97**	49.52	51.84	30.67	27.78	34.41	43.95	**35.76**	**40.35**	**38.05**	40.63
BERT$_E$-CDRG-Indep Training	51.01	54.56	54.33	42.52	39.28	36.98	40.23	33.41	30.56	32.05	41.49
BERT$_E$-CDRG-Merge Training	53.09	**57.96**	**54.39**	**40.85**	**42.96**	**38.83**	**45.66**	35.06	31.62	34.22	**43.46**

Table 2: Comparison results of different methods for Cross-Domain End-to-End ABSA based on Micro-F1.

Methods	S→R	L→R	D→R	R→S	L→S	D→S	R→L	S→L	R→D	S→D	Average
DP (Qiu et al., 2011)	37.63	37.63	37.63	19.74	19.74	19.74	19.79	19.79	21.82	21.82	25.53
Hier-Joint (Ding et al., 2017)	46.39	48.61	42.96	27.18	25.22	29.28	34.11	33.02	34.81	35.00	35.66
RNSCN (Wang and Pan, 2018)	48.89	52.19	50.39	30.41	31.21	35.50	47.23	34.03	**46.16**	32.41	40.84
AD-SAL (Li et al., 2019b)	52.05	56.12	51.55	39.02	38.26	36.11	45.01	35.99	43.76	**41.21**	43.91
BERT$_B$	54.29	46.74	44.63	22.31	30.66	33.33	37.02	36.88	32.03	38.06	37.60
BERT$_B$-UDA (Gong et al., 2020)	56.08	51.91	50.54	34.62	32.49	34.52	46.87	43.98	40.34	38.36	42.97
BERT$_B$-CDRG-Indep Training	53.79	55.13	50.07	41.74	44.14	37.1	40.18	33.22	30.78	34.97	42.11
BERT$_B$-CDRG-Merge Training	56.26	60.03	52.71	42.36	47.08	41.85	46.65	39.51	32.60	36.97	45.60
BERT$_E$	57.56	50.42	45.71	26.50	25.96	30.40	44.18	41.78	35.98	35.13	39.36
BERT$_E$-UDA (Gong et al., 2020)	59.07	55.24	56.40	34.21	30.68	38.25	54.00	**44.25**	42.40	40.83	45.53
BERT$_E$-CDRG-Indep Training	58.75	65.81	**59.61**	**50.68**	51.25	40.17	49.17	41.61	33.34	36.97	48.74
BERT$_E$-CDRG-Merge Training	**59.17**	**68.62**	58.85	47.61	**54.29**	**42.20**	**55.56**	41.77	35.43	36.53	**50.00**

Table 3: Comparison results of different methods for Cross-Domain Aspect Extraction based on Micro-F1.

For re-training the MLM task in our two-step approach for CDRG, we employ the Adam optimizer (Kingma and Ba, 2014) with a batch size of 32 and a learning rate of 3e-5. For the two training strategies in Section 4.2, we also adopt the Adam optimizer, in which the learning rate, the dropout rate and the batch size are set to 5e-5, 0.1, 32 after a grid search over the combinations of [2e-5, 8e-5], [0.1, 0.3] and {16, 32, 64}. These hyper-parameters are tuned on 10% randomly held-out training data from the source domain. All the experiments are run on a Nvidia GTX 1080 Ti GPU.

5.3 Main Results for Cross-Domain ABSA

We report the comparison results on the cross-domain E2E-ABSA task in Table 2, and make the following observations: (1) Comparing *Indep Training* with all the baseline methods, we can see that solely fine-tuning BERT$_E$ on our generated target-domain reviews can already outperform all the existing domain adaptation approaches on average, including our recent unified feature and instance-based adaptation (UDA) method. This demonstrates the usefulness of our CDRG-based approach. (2) By merging the generated target-domain reviews with the labeled source reviews, our *Merge Training* strategy further boosts the average performance of *Indep Training*, which outper-

forms the state-of-the-art *UDA* approach by an absolute improvement of 3.00% and 2.83% based on BERT$_B$ and BERT$_E$ respectively. All these observations verify the superiority of our CDRG-based approach over the previous feature and instance-based adaptation methods.

Similar to the results on cross-domain E2E-ABSA, Table 3 shows that our *Indep Training* strategy obtains indistinguishable performance compared with the state-of-the-art method UDA based on BERT$_B$, and achieves significantly better performance than UDA based on BERT$_E$. Moreover, our *Merge Training* strategy can consistently achieve the best average performance on the cross-domain AE task, which outperforms UDA by an absolute improvement of 2.63% and 4.47% based on BERT$_B$ and BERT$_E$ respectively. This demonstrates the general effectiveness of our CDRG-based domain adaptation methods.

5.4 Ablation Study of Our Two-Step Approach for CDRG

To investigate the effectiveness of our two-step approach for CDRG, we conduct the ablation study of our target-domain masked language model with constraints (TD-MLM-C) approach, and consider the following variants: (1) *BERT$_E$*: The pre-trained BERT$_E$ model without re-training MLM on the un-

	Source Domain: Laptop→Target Domain: Restaurant	
Source Review	1. The [screen graphics]_P and [clarity]_P, and [sharp ##ness]_P are great.	2. The [battery life]_P is great.
Target Review from BERT_E	The [laptop,]_P✗ and [sound]_P✗, and [the ##s]_P✗ are great.	The [touch screen]_P✗ is great.
Target Review from TD-MLM	The [food,]_P✗ and [service]_P✗, and [the prices]_P✗ are great.	The [su here]_P✗ is great.
Target Review from TD-MLM-C	The [pizza ##s]_P✓ and [atmosphere]_P✓, and [service staff]_P✓ are great.	The [su ##shi]_P✓ is great.
	Source Domain: Restaurant→ Target Domain: Laptop	
Source Review	3. The [food]_P is [flavor ##ful], [pl ##ent ##iful] and reasonably priced.	4. [Pizza]_P is terrific, as is [homemade pasta]_P.
Target Review from BERT_E	The [food]_P is [flavor ##ful]✗, [st ##ten ,]✗ and reasonably priced.	[It]_P✗ is terrific, as is [the price]_P✗.
Target Review from TD-MLM	The [keyboard]_P is [joy ##ful]✓, [easy ##yl ,]✗ and reasonably priced.	[It]_P✗ is terrific, as is [the screen]_P.
Target Review from TD-MLM-C	The [keyboard]_P is [joy ##ful]✓, [st ##yl ##ish]✓ and reasonably priced.	[Speed]_P✓ is terrific, as is [windows os]_P✓.

Table 4: Example target-domain reviews generated from BERT_E, TD-MLM, and TD-MLM-C. ✓ and ✗ indicate that the generated target-specific attributes are correct or incorrect. P and N denote the positive and negative sentiment. The blue and red colors refer to the domain-specific aspect terms and opinion terms, respectively. ## denotes that the original word is split into several sub-tokens by the tokenizer of BERT.

	Src&Tgt	BERT_E&Tgt	TD-MLM&Tgt	TD-MLM-C&Tgt
S→R	0.3496	0.2411	0.1621	**0.1477**
L→R	0.3155	0.2978	0.2456	**0.2072**
D→R	0.3504	0.3178	0.2109	**0.1650**
R→S	0.3330	0.2977	0.2516	**0.1654**
L→S	0.1966	0.1945	0.1742	**0.1518**
D→S	0.1804	0.1628	0.1167	**0.0941**
R→L	0.3135	0.2563	0.1589	**0.1080**
S→L	0.2767	0.1652	**0.1554**	0.1763
R→D	0.3476	0.2954	0.2026	**0.1433**
S→D	0.1765	0.0929	**0.0727**	0.0788
Average	0.2840	0.2322	0.1751	**0.1438**

Table 5: Maximum Mean Discrepancy (lower is better) between the target-domain test set and the review set generated from BERT_E, TD-MLM, and TD-MLM-C.

	Cross-Domain E2E-ABSA			Cross-Domain AE		
	BERT_E	TD-MLM	TD-MLM-C	BERT_E	TD-MLM	TD-MLM-C
S→R	46.42	47.32	**51.01**	53.04	52.87	**58.75**
L→R	52.40	**55.03**	54.56	59.22	62.57	**65.81**
D→R	44.90	50.43	**54.33**	48.88	54.21	**59.61**
R→S	22.24	34.67	**42.52**	24.58	39.41	**50.68**
L→S	23.98	35.84	**39.28**	26.31	41.87	**51.25**
D→S	22.79	31.44	**36.98**	25.09	34.00	**40.17**
R→L	33.89	39.21	**40.23**	39.73	48.49	**49.17**
S→L	29.88	30.78	**33.41**	34.61	35.64	**41.61**
R→D	30.32	29.36	**30.56**	32.91	32.80	**33.34**
S→D	25.67	26.69	**32.05**	28.07	28.68	**36.97**
Average	33.25	38.07	**41.49**	37.24	43.05	**48.74**

Table 6: Comparison results of using the labeled reviews generated from different methods as training instances for our Independent Training strategy.

labeled target data; (2) *TD-MLM*: re-training the BERT_E model with MLM on the unlabeled target data; (3) *TD-MLM-C*: adding target-specific attribute constraints into TD-MLM, as introduced in Section 4.1.2.

First, we propose to verify the closeness between the generated target-domain review set and the real test set from the target domain. Specifically, we employ BERT_E to obtain the sentence representation of each review in the two sets, and then compute the distance between the two sets with Maximum Mean Discrepancy (Gretton et al., 2012). Based on the results in Table 5, we can observe that all the three methods can consistently reduce the discrepancy between source and target domains. This shows our two-step approach is generally useful for domain adaptation. Moreover, it is clear that the distance between the review set from *TD-MLM-C* and the test set is the smallest, which implies the distribution of its generated reviews is closest to the distribution of the target domain.

Second, we further treat the generated reviews as training instances for our *Indep Training* strategy,

and compare their results on the cross-domain E2E-ABSA and AE tasks. From the results in Table 6, we can see that using the reviews from *TD-MLM-C* consistently achieves the best results, and outperforms the other methods with a significant margin.

5.5 Manual Evaluation on Generated Target-Domain Reviews

Since there is no ground-truth target review for each source review, we randomly select 200 source reviews from L→R and R→L transfer pairs, and manually evaluate the generated reviews in terms of coherence, label correctness, and domain-specific criteria. Based on our manual evaluation, we observe that our *TD-MLM-C* approach generates better target reviews than its two ablation systems for 152 source reviews; for the remaining 48 source reviews, we cannot determine the winning method, as all the three methods generate meaningful or unmeaningful target-domain reviews.

Table 4 shows four representative examples. We can find that *BERT_E* is generally insensitive to the target domain, which may still generate source-

specific terms (e.g., *touch screen* in S2 and *flavorful* in S3). *TD-MLM* can produce better reviews, as it tends to convert source-specific terms to the target domain. However, it still suffers from generating non-aspect or non-opinion terms (e.g., *the prices* in S1 and *It* in S4). In contrast, with the vocabulary constraints, *TD-MLM-C* can successfully convert all the source-specific attributes to target-specific attributes in the four cases.

All these observations verify the importance of TD-MLM and vocabulary constraints in our two-step approach for cross-domain review generation.

6 Conclusion

In this paper, we propose a new domain adaptation paradigm named Cross-Domain Review Generation (CDRG) with applications to the ABSA task. Specifically, we first propose a two-step approach to generate labeled target-domain reviews based on labeled source-domain reviews for CDRG, and then propose two training strategies to leverage the generated reviews for two cross-domain ABSA tasks. Experiments on four benchmark datasets demonstrate that our CDRG-based approaches significantly outperform existing methods for cross-domain E2E-ABSA and cross-domain AE tasks.

Acknowledgments

References

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*.

Minmin Chen, Zhixiang Xu, Kilian Q Weinberger, and Fei Sha. 2012. Marginalized denoising autoencoders for domain adaptation. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*.

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*.

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

Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. Adaptive recursive neural network for target-dependent twitter sentiment classification. In *Proceedings of ACL*.

Mark Dredze, Alex Kulesza, and Koby Crammer. 2010. Multi-domain learning by confidence-weighted parameter combination. *Machine Learning*, 79(1-2):123–149.

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.

Deepanway Ghosal, Devamanyu Hazarika, Abhinaba Roy, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2020. KinGDOM: Knowledge-Guided DOMain Adaptation for Sentiment Analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: a deep learning approach. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*.

Chenggong Gong, Jianfei Yu, and Rui Xia. 2020. Unified feature and instance based domain adaptation for end-to-end aspect-based sentiment analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2018. Adaptive semi-supervised learning for cross-domain sentiment classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*.

Jing Jiang and ChengXiang Zhai. 2007. Instance weighting for domain adaptation in nlp. In *Proceedings of the 45th annual meeting of the association of computational linguistics*.

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

Fangtao Li, Sinno Jialin Pan, Ou Jin, Qiang Yang, and Xiaoyan Zhu. 2012. Cross-domain co-extraction of sentiment and topic lexicons. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*.

Xin Li, Lidong Bing, Piji Li, and Wai Lam. 2019a. A unified model for opinion target extraction and target sentiment prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Xin Li, Lidong Bing, Piji Li, Wai Lam, and Zhimou Yang. 2018a. Aspect term extraction with history attention and selective transformation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*.

Zheng Li, Xin Li, Ying Wei, Lidong Bing, Yu Zhang, and Qiang Yang. 2019b. 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*.

Zheng Li, Ying Wei, Yu Zhang, and Qiang Yang. 2018b. Hierarchical attention transfer network for cross-domain sentiment classification. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Bing Liu. 2020. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge university press.

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*.

Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive attention networks for aspect-level sentiment classification. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*.

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

Margaret Mitchell, Jacqui Aguilar, Theresa Wilson, and Benjamin Van Durme. 2013. Open domain targeted sentiment. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*.

Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen. 2010. Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th international conference on World wide web*.

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

Maria Pontiki, Dimitrios Galanis, Harris 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)*.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Auresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*.

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.

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.

Duyu Tang, Bing Qin, and Ting Liu. 2016. Aspect level sentiment classification with deep memory network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Cigdem Toprak, Niklas Jakob, and Iryna Gurevych. 2010. Sentence and expression level annotation of opinions in user-generated discourse. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Shuai Wang, Sahisnu Mazumder, Bing Liu, Mianwei Zhou, and Yi Chang. 2018. Target-sensitive memory networks for aspect sentiment classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

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*.

Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2016a. Recursive neural conditional random fields for aspect-based sentiment analysis. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2017. Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016b. Attention-based lstm for aspect-level sentiment classification. In *EMNLP 2016: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Rui Xia, Jianfei Yu, Feng Xu, and Shumei Wang. 2014. Instance-based domain adaptation in nlp via in-target-domain logistic approximation. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Hu Xu, Bing Liu, Lei Shu, and S Yu Philip. 2018. Double embeddings and cnn-based sequence labeling for aspect extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

Hu Xu, Bing Liu, Lei Shu, and S Yu Philip. 2019. Bert post-training for review reading comprehension and aspect-based sentiment analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Min Yang, Wenting Tu, Jingxuan Wang, Fei Xu, and Xiaojun Chen. 2017. Attention based lstm for target dependent sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Hai Ye, Qingyu Tan, Ruidan He, Juntao Li, Hwee Tou Ng, and Lidong Bing. 2020. Feature adaptation of pre-trained language models across languages and domains with robust self-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Jianfei Yu and Jing Jiang. 2016. Learning sentence embeddings with auxiliary tasks for cross-domain sentiment classification. In *EMNLP 2016: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Jianfei Yu and Jing Jiang. 2017. Leveraging auxiliary tasks for document-level cross-domain sentiment classification. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*.

Meishan Zhang, Yue Zhang, and Duy-Tin Vo. 2015. Neural networks for open domain targeted sentiment. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Fuzhen Zhuang, Xiaohu Cheng, Ping Luo, Sinno Jialin Pan, and Qing He. 2015. Supervised representation learning: Transfer learning with deep autoencoders. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Yftah Ziser and Roi Reichart. 2018. 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*.