Transferable End-to-End Aspect-based Sentiment Analysis with Selective Adversarial Learning

Zheng Li[1], Xin Li[2], Ying Wei[3], Lidong Bing[4], Yu Zhang[5], Qiang Yang[1]

[1]The Hong Kong University of Science and Technology

[2]The Chinese University of Hong Kong

[3]Tencent AI Lab

[4]Alibaba DAMO Academy

[5]Southern University of Science and Technology

Outline

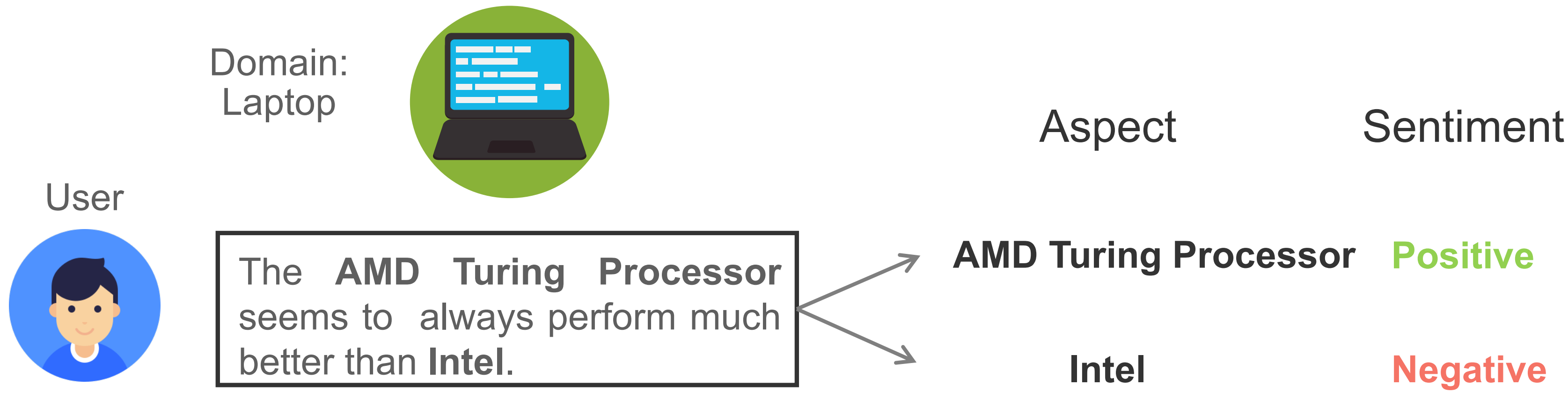
- Background & Motivation
  - E2E-ABSA
  - Transferable E2E-ABSA
- Method
- Experiments
- Future Works

Outline

- Background & Motivation
  - E2E-ABSA
  - Transferable E2E-ABSA
- Method
- Experiments
- Future Works

End-to-End Aspect-based Sentiment Analysis

E2E-ABSA: joint extraction of aspects and their sentiments from user reviews.



Two subtasks:

- Aspect detection (**AD**): extract the aspect terms from user reviews.
- Aspect sentiment (**AS**) classification: Given a review sentence and an aspect term, predict the sentiment towards the aspect.

Unified Formulation (single domain)

Formulation: coupling two subtasks as a unified sequence labeling problem.

- **Unified tag** = **aspect boundary tag** {B, I, E, O, S} + **sentiment tag** {POS, NEG, NEU}
- **NER tag** = **entity boundary tag** {B, I, E, O, S} + **entity type tag** {PER, ORG, LOC, …} (***Similar***)

Input	The	AMD	Turin	Processor	seems	to	always	perform	much	better	than	Intel	.
Joint	O	B	I	E	O	O	O	O	O	O	O	S	O
	O	POS	POS	POS	O	O	O	O	O	O	O	NEG	O
Unified	O	B-POS	I-POS	E-POS	O	O	O	O	O	O	O	S-NEG	O

Li et al., 2019 A unified model for opinion target extraction and target sentiment prediction AAAI

Pros:

- End-to-end manner
- Alleviate accumulated errors across two highly-correlative sub-tasks.

Cons:

- Lack of sufficient labeled data in a wide range of domains.
- Manual labeling for sequence data is expensive and time-consuming.

Transferable E2E-ABSA

Source domain
(Train): Laptop

User

The **[AMD Turing Processor]_POS** seems to always perform much better than **[Intel]_NEG**.

Cross-domain:

- leverage knowledge from a labeled source domain to improve the **sequence learning** (unified tag prediction) in an unlabeled target domain.

User

Target domain:
(**Test**) : Restaurant

Great salmon but the waitress is so rude.

Aspect

Sentiment

Unified Formulation (cross domain)

Input: a sequence of words $\mathbf{x} = \{w_1, w_2, ..., w_n\}$

Output: a unified tag sequence $\mathbf{y} = \{y_1, y_2, ..., y_n\}$

Source

$$D_s = \{(\mathbf{x}_s^i, \mathbf{y}_s^i)\}_{i=1}^{N_s}$$

Target

$$D_t = \{(\mathbf{x}_t^j)\}_{j=1}^{N_t}$$

Sequence Transfer Learning (unsupervised)

Challenges

What to transfer?

- There exists a **large domain shift** between domains since aspect terms in different domains are usually disjoint.

  e.g., “salmon” in the Restaurant domain and “mouse” in the Laptop domain.

How to transfer?

- Unlike domain adaptation in traditional sentiment classification that learns shared sentence or document representations, we need to learn **fine-grained** (word-level) representations to be domain-invariant for **sequence prediction**.

What to transfer?

Prior work: highly depends on common syntactic relations between aspect and opinion words

- manually-designed rules
- external linguistic resources (dependency parsers)

I love tuna sandwich very much.

I love the design of iPhone 7

RuleID	Rule	Example
R1	$\mathtt{O} \xrightarrow{amod} \mathtt{T}$	They have nice dessert. (nice $\xrightarrow{amod}$ *dessert*)
R2	$\mathtt{T} \xrightarrow{nsubj} \mathtt{O}$	Its camera is great. (*camera* $\xrightarrow{nsubj}$ great)
R3	$\mathtt{T} \xrightarrow{dobj} \mathtt{O}$	I love their fries. (*fries* $\xrightarrow{dobj}$ love)
R4	$\mathtt{T} \xrightarrow{nsubj} \mathtt{H} \xleftarrow{amod} \mathtt{O}$	iPhone is the best cellphone. (*iPhone* $\xrightarrow{nsubj}$ phone $\xleftarrow{amod}$ best)
ER1	$\mathtt{W} \xrightarrow{amod} \mathtt{T}$	I like Indian food. (*Indian* $\xrightarrow{amod}$ food)
ER2	$\mathtt{W} \xrightarrow{nn} \mathtt{T}$	Their spring roll is great. (*spring* $\xrightarrow{nn}$ roll)
ER3	$\mathtt{W}_2 \xrightarrow{pobj} \mathtt{W}_1 \xrightarrow{prep} \mathtt{T}$	I like the design of iPhone. (*iPhone* $\xrightarrow{pobj}$ *of* $\xrightarrow{prep}$ design)

Ding et al., 2017. Recurrent neural networks with auxiliary labels for cross-domain opinion target extraction. AAAI

Ours: automatically capture the latent relations among aspect and opinion words as transferable knowledge.

How to transfer?

straightforward solution: apply domain adaption methods to align all words within the sentence. (**no significant improvements**).

Reason: Only a small number of words are informative words that are not tagged with "O".

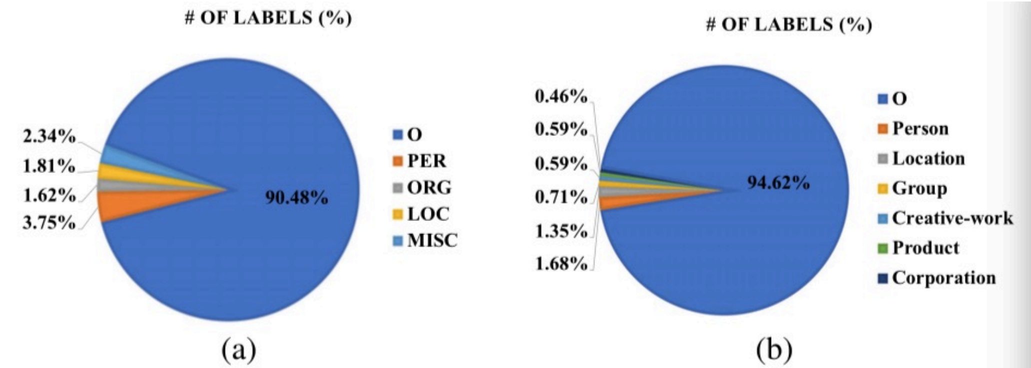


Fig. 2. Label statistics on CoNLL-2002 Dutch NER and WNUT-2017 English Twitter NER train set. The label type "O" is dominant in different data sets. (a) CoNLL-2002 Dutch. (b) WNUT-2017 English Twitter.

TABLE I
F1 SCORE OF BILSTM-CNNS-CRF MODEL ON WNUT-2017 DEVELOPMENT SET

Label Type	Precision (%)	Recall (%)	F1 Score (%)
O	95.56	99.46	97.47
Person	78.75	49.23	60.59
Location	51.02	46.73	48.78
Group	19.23	7.81	11.11
Creative-work	32.94	11.76	17.34
Product	33.33	7.69	12.50
Corporation	18.18	21.74	19.80

Zhou et al., Roseq: Robust sequence labeling. TNNLS 2019

Ours: selectively align the informative words within the sentence.

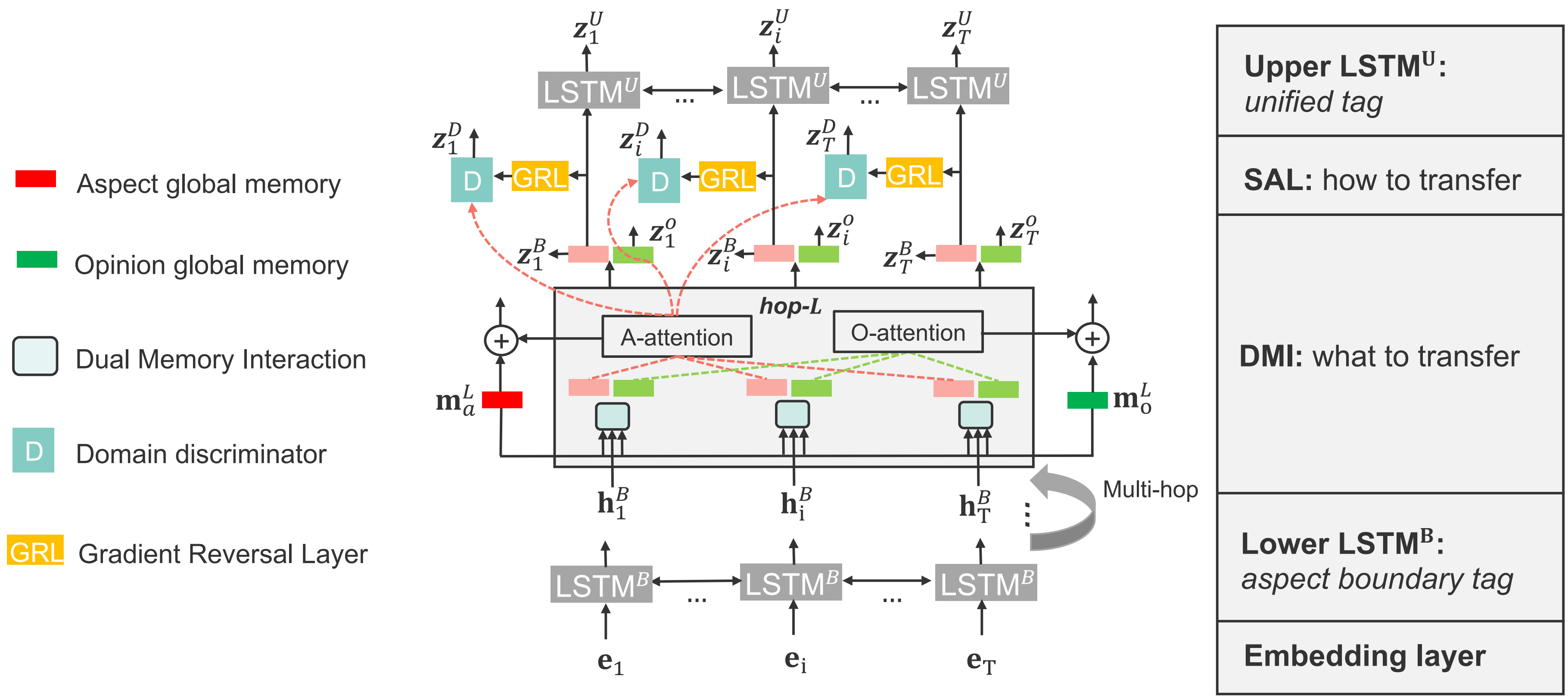
Outline

Method

Experiments

Future Works

Framework


Aspect global memory
Opinion global memory
Dual Memory Interaction
Domain discriminator
Gradient Reversal Layer
LSTM^U
LSTM^U
LSTM^U
GRL
GRL
GRL
D
D
D
hop-L
A-attention
O-attention
Multi-hop
LSTM^B
LSTM^B
LSTM^B
Upper LSTM^U: unified tag
SAL: how to transfer
DMI: what to transfer
Lower LSTM^B: aspect boundary tag
Embedding layer

Base Model

Two stacked Bi-LSTMs:

- Aspect boundary information can be used as the guidance

$$
\begin{aligned}
h_i^{\mathcal{B}} &= [\overrightarrow{\mathrm{LSTM}}^{\mathcal{B}}(e_i); \overleftarrow{\mathrm{LSTM}}^{\mathcal{B}}(e_i)], \\
h_i^{\mathcal{U}} &= [\overrightarrow{\mathrm{LSTM}}^{\mathcal{U}}(h_i^{\mathcal{B}}); \overleftarrow{\mathrm{LSTM}}^{\mathcal{U}}(h_i^{\mathcal{B}})].
\end{aligned}
$$

Low-level AD $\quad z_i^{\mathcal{B}} = p(y_i^{\mathcal{B}} | h_i^{\mathcal{B}}) = \mathrm{Softmax}(W_{\mathcal{B}} h_i^{\mathcal{B}} + b_{\mathcal{B}}).$

High-level ADS $\quad z_i^{\mathcal{U}} = p(y_i^{\mathcal{U}} | h_i^{\mathcal{U}}) = \mathrm{Softmax}(W_{\mathcal{U}} h_i^{\mathcal{U}} + b_{\mathcal{U}}).$

Primary loss:

$$
\mathcal{L}_{\mathcal{M}} = \sum_{D_s} \sum_{\mathcal{Q} \in \{\mathcal{B}, \mathcal{U}\}} \sum_{i=1}^{T} \ell(z_i^{\mathcal{Q}}, y_i^{\mathcal{Q}}).
$$

AD: Aspect Detection (aspect boundary tags)

ADS: Aspect Detection and Sentiment Classification (unified tags)

What to transfer: Global-Local Memory Interaction (GLMI)

GLMI: basic operation for computing the correlations between two objects (local & global memory).

$$\text{GLMI: } f(\overbrace{\mathbf{h}_i^{\mathcal{B}}}^{\text{Local Memory: LSTM hidden states}}, \underbrace{\mathbf{m}}_{\text{Global Memory: commonly-used in memory networks}}; \mathbf{\Theta}, \mathbf{G})$$

1) Residual transformation

$$\tilde{\mathbf{h}}_i^{\mathcal{B}} = \mathbf{h}_i^{\mathcal{B}} + \mathrm{ReLU}(\mathbf{W}[\mathbf{h}_\mathbf{i}^{\mathcal{B}} : \mathbf{m}] + \mathbf{b}),$$

2) Multi-dimensional Bilinear Transformation

$$\mathbf{r}_i = \mathbf{m}^T \underbrace{\mathbf{G}}_{\mathbf{G} \in \mathbb{R}^{K \times \dim_h^{\mathcal{B}} \times \dim_h^{\mathcal{B}}} \text{ models K kinds of latent relations}} \tilde{\mathbf{h}}_i^{\mathcal{B}}$$

What to transfer: Dual Memory Interaction (DMI)

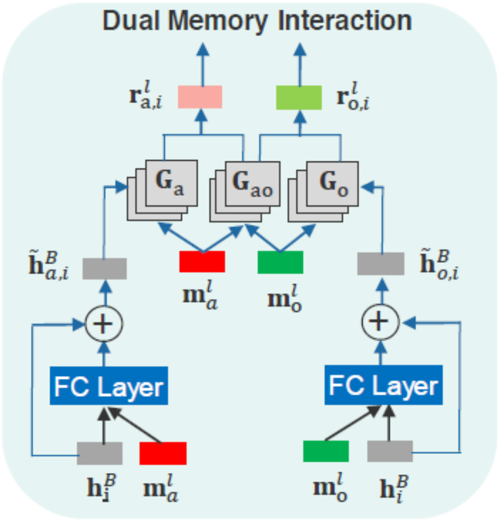
DMI: models the correlations between **a**spect and **o**pinions for aspect and opinion co-detection.

aspect $$\mathbf{r}^l_{a,i} = [\overbrace{f(\mathbf{h}_i^{\mathcal{B}}, \mathbf{m}_a^l; \boldsymbol{\Theta}_a, \mathbf{G}_a)}^{\textbf{a \& a}} : \overbrace{f(\mathbf{h}_i^{\mathcal{B}}, \mathbf{m}_o^l; \boldsymbol{\Theta}_o, \mathbf{G}_{ao})}^{\textbf{a \& o}}],$$

opinion $$\mathbf{r}^l_{o,i} = [\underbrace{f(\mathbf{h}_i^{\mathcal{B}}, \mathbf{m}_o^l; \boldsymbol{\Theta}_o, \mathbf{G}_o)}_{\textbf{o \& o}} : \underbrace{f(\mathbf{h}_i^{\mathcal{B}}, \mathbf{m}_a^l; \boldsymbol{\Theta}_a, \mathbf{G}_{ao}^T)}_{\textbf{o \& a}}],$$

opinion detection is used as a auxiliary task to link the different aspect across domains.

$$\mathcal{L}_{\mathcal{O}} = \sum_{D_s \cup D_t} \ell(\mathbf{z}_i^{\mathcal{O}}, \mathbf{y}_i^{\mathcal{O}})$$



global aspect memory

global opinion memory

A-attention (aspect) & O-attention (opinion)

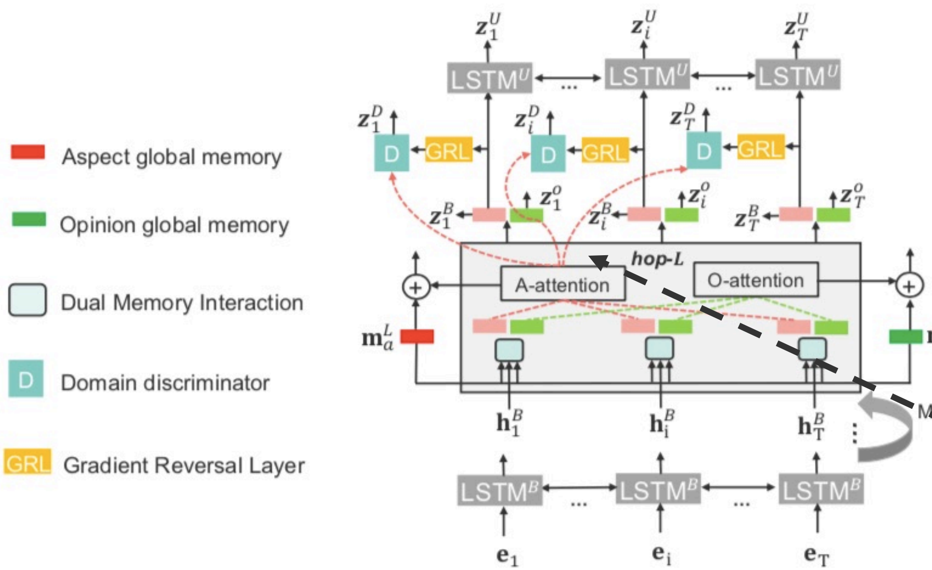
$$\alpha^l_{p,i} = \frac{\exp(\mathbf{u}_p \mathbf{r}^l_{p,i})}{\sum_{j=1}^T \exp(\mathbf{u}_p \mathbf{r}^l_{p,j})}. \qquad p \in \{a, o\}$$

How to transfer: Selective Adversarial Learning (SAL)

Gradient Reversal Layer: (Ganin et al., 2016)

$$R_\lambda(\mathbf{x}) = \mathbf{x}$$

$$\frac{\partial R_\lambda(\mathbf{x})}{\partial \mathbf{x}} = -\lambda I,$$



Domain discriminator

$$\mathbf{z}_i^{\mathcal{D}} = \mathbf{p}(\mathbf{y}_i^{\mathcal{D}} | \mathbf{r}_{a,i}^L) = \mathrm{Softmax}(\mathbf{W}_{\mathcal{D}} R_\lambda(\mathbf{r}_{a,i}^L) + \mathbf{b}_{\mathcal{D}})$$

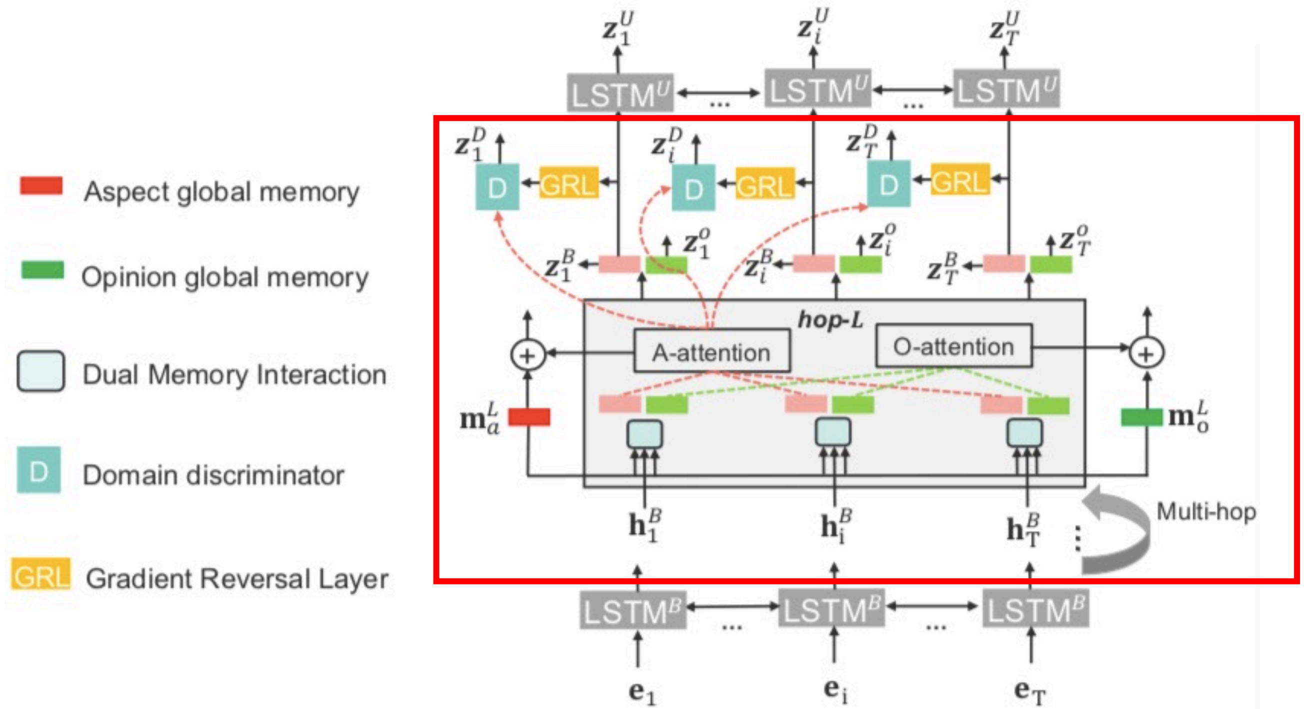
Selective adversarial loss:

$$\mathcal{L}_{\mathcal{D}} = \sum_{D_s \cup D_t} \sum_{i=1}^{T} \underbrace{\alpha_{a,i}^L}_{\text{A-attention dynamically controls the selectivity}} \ell(\mathbf{z}_i^{\mathcal{D}}, \mathbf{y}_i^{\mathcal{D}}).$$

A-attention dynamically controls the selectivity

How to transfer: Selective Adversarial Learning (SAL)

- Why do we choose low-level neural layer features (e.g., SAL on the low-level AD task) for transfer?



Existing studies (Yosinski et al., 2014; Mou et al., 2016) have already shown some evidence that low-level neural layer features (i.e., low-level task) are more easily transferred to different tasks or domains.

Training Strategy

Joint training: (unstable, too many objectives)

$$E = \mathcal{L}_{\mathcal{M}} + \rho \mathcal{L}_{\mathcal{O}} + \gamma \mathcal{L}_{\mathcal{D}}$$

Alternating training: (more unstable, two-stage optimization)

$$\mathcal{L}_{\mathcal{M}} + \rho \mathcal{L}_{\mathcal{O}}$$

$$\mathcal{L}_{\mathcal{D}}$$

iterative

$$(\hat{\boldsymbol{\theta}}_f^{(1)}, \hat{\boldsymbol{\theta}}_w) = \arg \min_{\boldsymbol{\theta}_f, \boldsymbol{\theta}_w} \mathcal{L}_{\mathcal{M}} + \rho \mathcal{L}_{\mathcal{O}}$$

discriminative stage

$$(\hat{\boldsymbol{\theta}}_f^{(2)}, \hat{\boldsymbol{\theta}}_d) = \arg \min_{\boldsymbol{\theta}_d} \max_{\boldsymbol{\theta}_f} \mathcal{L}_{\mathcal{D}}.$$

domain-invariant stage

$\boldsymbol{\theta}_f, \ \boldsymbol{\theta}_w, \ \boldsymbol{\theta}_d$ The parameters for feature learning of each word, word predictions for AD, ADS and opinion detection tasks, and domain classification, respectively.

Outline

Method

Experiments

Future Works

Experiment Setup

Datasets

- **L**: Laptop domain (Pontiki et al., 2014)
- **R**: Restaurant domain (Pontiki et al., 2014,2015,2016)
- **D**: Device domain (Hu and Liu, 2004)
- **S**: Service domain (Toprak et al., 2010)

L and **R** are from SemEval ABSA challenge 2014, 2015, 2016

Dataset	Domain	Sentences	Training	Testing
$\mathbb{L}$	Laptop	1,869	1,458	411
$\mathbb{R}$	Restaurant	3,900	2,481	1,419
$\mathbb{D}$	Device	1,437	954	483
$\mathbb{S}$	Service	2,153	1,433	720

Experiment Setup

Setting

4 different domains, 10 transfer pairs (without two easy pairs L→D, D→L)

For each pair e.g., **source** A-> **target** B:

Training: Labeled training data from A, unlabeled training data from B

Validation: testing data from A

Testing: testing data from B

Baselines

- **TCRF**: (Jakob and Gurevych, 2010): Transferable CRF
- **RAP**: (Li et al., 2012): cross-domain Relational Adaptive Bootstrapping
- **Hier-Joint**: (Ding et al., 2017): RNN with manually designed rule-based auxiliary tasks based on common syntactic relations
- **RNSCN**: (Wang and Pan, 2018): a recursive neural structural correspondence network

Extended versions:

- **Hier-Joint+ & RNSCN+**: original version with the proposed stacking architecture

Main Results

Transfer Pair	TCRF		RAP		Hier-Joint		Hier-Joint$^+$		RNSCN		RNSCN$^+$		**Ours**	
	AD	ADS	AD	ADS	AD	ADS	AD	ADS	AD	ADS	AD	ADS	AD	ADS
$\mathbb{S}\rightarrow\mathbb{R}$	-	14.84	-	25.41	-	32.81	46.39	31.10	-	30.56	48.89	33.21	**52.05**	**41.03**
$\mathbb{L}\rightarrow\mathbb{R}$	-	16.06	-	31.05	-	31.90	48.61	33.54	-	31.85	52.19	35.65	**56.12**	**43.04**
$\mathbb{D}\rightarrow\mathbb{R}$	-	17.05	-	28.37	-	30.03	42.96	32.87	-	31.41	50.39	34.60	**51.55**	**41.01**
$\mathbb{R}\rightarrow\mathbb{S}$	-	15.20	-	13.17	-	15.20	27.18	15.56	-	23.31	30.41	20.04	**39.02**	**28.01**
$\mathbb{L}\rightarrow\mathbb{S}$	-	12.34	-	13.72	-	15.33	25.22	13.90	-	16.73	31.21	16.59	**38.26**	**27.20**
$\mathbb{D}\rightarrow\mathbb{S}$	-	13.49	-	16.80	-	18.74	29.28	19.04	-	18.93	35.50	20.03	**36.11**	**26.62**
$\mathbb{R}\rightarrow\mathbb{L}$	-	14.59	-	15.69	-	19.17	34.11	20.72	-	25.54	**47.23**	26.63	45.01	**34.13**
$\mathbb{S}\rightarrow\mathbb{L}$	-	9.56	-	12.38	-	21.80	33.02	22.65	-	19.15	34.03	18.87	**35.99**	**27.04**
$\mathbb{R}\rightarrow\mathbb{D}$	-	19.84	-	17.50	-	22.91	34.81	24.53	-	32.43	**46.16**	33.26	43.76	**35.44**
$\mathbb{S}\rightarrow\mathbb{D}$	-	13.43	-	15.74	-	20.04	35.00	23.24	-	19.98	32.41	22.00	**41.21**	**33.56**
Average	-	14.64	-	18.98	-	22.79	35.66	23.72	-	24.99	40.84	26.09	**43.91**†	**33.71**†
(Δ)	-	(19.07)	-	(14.73)	-	(10.92)	(8.25)	(9.99)	-	(8.72)	(3.07)	(7.62)	-	-

Table 2: Main results (%). Δ refers to the improvements of the full model over baseline methods. The marker $\dagger$ means that our model significantly outperforms the best baseline **RNSCN**$^+$ with p-value < 0.01.

Ablation Study

Ablation Variants

What to transfer

- **Base Model (SO / TO):** two stacked Bi-LSTMs. SO (Source Only) and TO (Target Only). We usually refer to them as a lower bound and a upper bound, respectively.
- **Base Model + DMI:** two stacked Bi-LSTMs with a DMI between them.

How to transfer

- **AD-AL:** pure adversarial learning (removing the selective weight from the adversarial loss) for the low-level AD task .
- **ADS-SAL:** selective adversarial learning on each word representations for the high-level ADS task.
- **AD-SAL (Full model):** selective adversarial learning for the low-level AD task .

Note: The backbones of the **AD-AL ADS-SAL** and **AD-SAL** are all based on the **Base Model +DMI**

No DMI v.s. DMI

No DMI DMI

Transfer Pair	Lower bound		Ablation Models						Full Model		Upper bound	
	Base Model (SO)		Base Model+DMI		AD-AL		ADS-SAL		AD-SAL		Base Model (TO)	
	AD	ADS	AD	ADS	AD	ADS	AD	ADS	AD	ADS	AD	ADS
$\mathbb{S}\to\mathbb{R}$	30.32	19.74	45.68	37.10	48.28	37.65	51.29	41.03	**52.05**	**41.03**	81.84	67.26
$\mathbb{L}\to\mathbb{R}$	33.99	28.34	46.25	36.49	51.79	38.63	55.50	42.00	**56.12**	**43.04**		
$\mathbb{D}\to\mathbb{R}$	31.59	27.25	46.56	36.89	46.39	37.34	46.43	38.35	**51.55**	**41.01**		
$\mathbb{R}\to\mathbb{S}$	15.63	8.61	21.88	16.85	25.13	18.61	37.11	25.84	**39.02**	**28.01**	68.28	41.12
$\mathbb{L}\to\mathbb{S}$	22.45	16.07	28.67	21.53	28.18	20.74	30.35	23.73	**38.26**	**27.20**		
$\mathbb{D}\to\mathbb{S}$	16.79	9.49	31.91	22.14	32.88	24.89	32.51	21.45	**36.11**	**26.62**		
$\mathbb{R}\to\mathbb{L}$	38.45	23.40	42.27	30.52	40.52	28.77	44.56	33.34	**45.01**	**34.13**	75.95	52.62
$\mathbb{S}\to\mathbb{L}$	24.69	14.48	36.38	27.48	32.96	25.16	33.87	24.22	**35.99**	**27.04**		
$\mathbb{R}\to\mathbb{D}$	34.87	25.79	36.90	27.71	41.61	31.88	**43.97**	34.50	43.76	**35.44**	70.37	57.62
$\mathbb{S}\to\mathbb{D}$	27.73	17.73	38.03	31.21	39.54	32.28	40.40	33.26	**41.21**	**33.56**		
Average	27.65	19.09	37.45	28.79	38.73	29.60	41.60	31.77	**43.91**†	**33.71**†	74.11	54.66
(Δ)	(16.26)	(14.62)	(6.46)	(4.92)	(5.18)	(4.11)	(2.31)	(1.94)	-	-	-	-

Table 3: Ablation results (%). Δ refers to the improvements of the full model over ablation methods. The marker † means that the full model significantly outperforms the best ablation model **ADS-SAL** with p-value < 0.01.

No SAL v.s. SAL

No SAL

SAL

Transfer Pair	**Lower bound**		**Ablation Models**						**Full Model**		**Upper bound**	
	Base Model (SO)		Base Model+DMI		AD-AL		ADS-SAL		AD-SAL		Base Model (TO)	
	AD	ADS	AD	ADS	AD	ADS	AD	ADS	AD	ADS	AD	ADS
$\mathbb{S} \to \mathbb{R}$	30.32	19.74	45.68	37.10	48.28	37.65	51.29	41.03	**52.05**	**41.03**		
$\mathbb{L} \to \mathbb{R}$	33.99	28.34	46.25	36.49	51.79	38.63	55.50	42.00	**56.12**	**43.04**	81.84	67.26
$\mathbb{D} \to \mathbb{R}$	31.59	27.25	46.56	36.89	46.39	37.34	46.43	38.35	**51.55**	**41.01**		
$\mathbb{R} \to \mathbb{S}$	15.63	8.61	21.88	16.85	25.13	18.61	37.11	25.84	**39.02**	**28.01**		
$\mathbb{L} \to \mathbb{S}$	22.45	16.07	28.67	21.53	28.18	20.74	30.35	23.73	**38.26**	**27.20**	68.28	41.12
$\mathbb{D} \to \mathbb{S}$	16.79	9.49	31.91	22.14	32.88	24.89	32.51	21.45	**36.11**	**26.62**		
$\mathbb{R} \to \mathbb{L}$	38.45	23.40	42.27	30.52	40.52	28.77	44.56	33.34	**45.01**	**34.13**	75.95	52.62
$\mathbb{S} \to \mathbb{L}$	24.69	14.48	36.38	27.48	32.96	25.16	33.87	24.22	**35.99**	**27.04**		
$\mathbb{R} \to \mathbb{D}$	34.87	25.79	36.90	27.71	41.61	31.88	**43.97**	34.50	43.76	**35.44**	70.37	57.62
$\mathbb{S} \to \mathbb{D}$	27.73	17.73	38.03	31.21	39.54	32.28	40.40	33.26	**41.21**	**33.56**		
Average	27.65	19.09	37.45	28.79	38.73	29.60	41.60	31.77	**43.91**†	**33.71**†	74.11	54.66
(Δ)	(16.26)	(14.62)	(6.46)	(4.92)	(5.18)	(4.11)	(2.31)	(1.94)	-	-	-	-

Table 3: Ablation results (%). Δ refers to the improvements of the full model over ablation methods. The marker † means that the full model significantly outperforms the best ablation model **ADS-SAL** with p-value < 0.01.

No Selectivity v.s. Selectivity

No Selectivity

Selectivity

Transfer Pair	Lower bound: Base Model (SO)		Ablation Models: Base Model+DMI		Ablation Models: AD-AL		Ablation Models: ADS-SAL		Full Model: AD-SAL		Upper bound: Base Model (TO)	
	AD	ADS	AD	ADS	AD	ADS	AD	ADS	AD	ADS	AD	ADS
$\mathbb{S}\to\mathbb{R}$	30.32	19.74	45.68	37.10	48.28	37.65	51.29	41.03	**52.05**	**41.03**		
$\mathbb{L}\to\mathbb{R}$	33.99	28.34	46.25	36.49	51.79	38.63	55.50	42.00	**56.12**	**43.04**	81.84	67.26
$\mathbb{D}\to\mathbb{R}$	31.59	27.25	46.56	36.89	46.39	37.34	46.43	38.35	**51.55**	**41.01**		
$\mathbb{R}\to\mathbb{S}$	15.63	8.61	21.88	16.85	25.13	18.61	37.11	25.84	**39.02**	**28.01**		
$\mathbb{L}\to\mathbb{S}$	22.45	16.07	28.67	21.53	28.18	20.74	30.35	23.73	**38.26**	**27.20**	68.28	41.12
$\mathbb{D}\to\mathbb{S}$	16.79	9.49	31.91	22.14	32.88	24.89	32.51	21.45	**36.11**	**26.62**		
$\mathbb{R}\to\mathbb{L}$	38.45	23.40	42.27	30.52	40.52	28.77	44.56	33.34	**45.01**	**34.13**	75.95	52.62
$\mathbb{S}\to\mathbb{L}$	24.69	14.48	36.38	27.48	32.96	25.16	33.87	24.22	**35.99**	**27.04**		
$\mathbb{R}\to\mathbb{D}$	34.87	25.79	36.90	27.71	41.61	31.88	**43.97**	34.50	43.76	**35.44**	70.37	57.62
$\mathbb{S}\to\mathbb{D}$	27.73	17.73	38.03	31.21	39.54	32.28	40.40	33.26	**41.21**	**33.56**		
Average	27.65	19.09	37.45	28.79	38.73	29.60	41.60	31.77	**43.91**†	**33.71**†	74.11	54.66
(Δ)	(16.26)	(14.62)	(6.46)	(4.92)	(5.18)	(4.11)	(2.31)	(1.94)	-	-	-	-

Table 3: Ablation results (%). Δ refers to the improvements of the full model over ablation methods. The marker † means that the full model significantly outperforms the best ablation model **ADS-SAL** with p-value < 0.01.

Low-level v.s. High-level

High-level Low-level

Transfer Pair	Lower bound		Ablation Models						Full Model		Upper bound	
	Base Model (SO)		Base Model+DMI		AD-AL		ADS-SAL		AD-SAL		Base Model (TO)	
	AD	ADS	AD	ADS	AD	ADS	AD	ADS	AD	ADS	AD	ADS
$\mathbb{S} \to \mathbb{R}$	30.32	19.74	45.68	37.10	48.28	37.65	51.29	41.03	**52.05**	**41.03**		
$\mathbb{L} \to \mathbb{R}$	33.99	28.34	46.25	36.49	51.79	38.63	55.50	42.00	**56.12**	**43.04**	81.84	67.26
$\mathbb{D} \to \mathbb{R}$	31.59	27.25	46.56	36.89	46.39	37.34	46.43	38.35	**51.55**	**41.01**		
$\mathbb{R} \to \mathbb{S}$	15.63	8.61	21.88	16.85	25.13	18.61	37.11	25.84	**39.02**	**28.01**		
$\mathbb{L} \to \mathbb{S}$	22.45	16.07	28.67	21.53	28.18	20.74	30.35	23.73	**38.26**	**27.20**	68.28	41.12
$\mathbb{D} \to \mathbb{S}$	16.79	9.49	31.91	22.14	32.88	24.89	32.51	21.45	**36.11**	**26.62**		
$\mathbb{R} \to \mathbb{L}$	38.45	23.40	42.27	30.52	40.52	28.77	44.56	33.34	**45.01**	**34.13**	75.95	52.62
$\mathbb{S} \to \mathbb{L}$	24.69	14.48	36.38	27.48	32.96	25.16	33.87	24.22	**35.99**	**27.04**		
$\mathbb{R} \to \mathbb{D}$	34.87	25.79	36.90	27.71	41.61	31.88	**43.97**	34.50	43.76	**35.44**	70.37	57.62
$\mathbb{S} \to \mathbb{D}$	27.73	17.73	38.03	31.21	39.54	32.28	40.40	33.26	**41.21**	**33.56**		
Average	27.65	19.09	37.45	28.79	38.73	29.60	41.60	31.77	**43.91**†	**33.71**†	74.11	54.66
(Δ)	(16.26)	(14.62)	(6.46)	(4.92)	(5.18)	(4.11)	(2.31)	(1.94)	-	-	-	-

Table 3: Ablation results (%). Δ refers to the improvements of the full model over ablation methods. The marker † means that the full model significantly outperforms the best ablation model **ADS-SAL** with p-value < 0.01.

Case studies

No Adaptation | Adversarial | Selective Adversarial

Input: (Target domain $\mathbb{L}$)	Base model+DMI		AD-AL		AD-SAL	
	AD	ADS	AD	ADS	AD	ADS
1. This laptop has only 2 [*usb ports*]$_{\text{NEG}}$, and they are both on the same side .	*ports*(✗), *side* (✗)	NONE(✗)	NONE(✗)	NONE(✗)	*usb ports*	[*usb ports*]$_{\text{NEG}}$
2. It is very easy to integrate [*bluetooth devices*]$_{\text{POS}}$, and [*usb devices*]$_{\text{POS}}$ are recognized almost instantly .	*devices* (✗), *devices* (✗)	[*devices*]$_{\text{POS}}$ (✗), [*devices*]$_{\text{POS}}$ (✗)	NONE(✗)	NONE(✗)	*bluetooth devices*, *usb devices*	[*bluetooth devices*]$_{\text{POS}}$, [*usb devices*]$_{\text{POS}}$
3. I also wanted [*windows 7*]$_{\text{POS}}$, which this one has .	NONE(✗)	NONE(✗)	NONE(✗)	NONE(✗)	*windows 7*	[*windows 7*]$_{\text{POS}}$
4. The [*speed*]$_{\text{POS}}$, the [*simplicity*]$_{\text{POS}}$, the [*design*]$_{\text{POS}}$ it is lightyears ahead of any pc i have ever owned .	*speed*, *design*	[*speed*]$_{\text{POS}}$, [*design*]$_{\text{POS}}$	*speed*, *design*, *pc* (✗)	[*speed*]$_{\text{POS}}$, [*design*]$_{\text{POS}}$, [*pc*]$_{\text{POS}}$ (✗)	*speed*, *design*, *simplicity*	[*speed*]$_{\text{POS}}$, [*design*]$_{\text{POS}}$, [*simplicity*]$_{\text{POS}}$
6. The [*battery life*]$_{\text{POS}}$ is excellent , the [*display*]$_{\text{POS}}$ is excellent and [*downloading apps*]$_{\text{POS}}$ is a breeze .	*battery* (✗), *display*, *apps* (✗)	[*battery*]$_{\text{POS}}$ (✗), [*display*]$_{\text{POS}}$, [*apps*]$_{\text{POS}}$ (✗)	*battery* (✗), *display*, *apps* (✗)	[*battery*]$_{\text{POS}}$ (✗), [*display*]$_{\text{POS}}$, [*apps*]$_{\text{POS}}$ (✗)	*battery life*, *display*, *downloading apps*	[*battery life*]$_{\text{POS}}$, [*display*]$_{\text{POS}}$, [*downloading apps*]$_{\text{POS}}$

Table 4: Case analysis for the $\mathbb{R}\rightarrow\mathbb{L}$ pair. Note that we only show the sentiment part of the unified labels (i.e., POS, NEG, and NEU) and use brackets to indicate the boundary. The marker ✗ denotes an incorrect prediction.

Outline

Method

Experiments

Future Works

Future Works

- Potentially, extend the proposed SAL method to other domain adaptation methods.
- Apply the SAL on more general sequence labeling tasks including NER, POS, Chunking and so on.

Thank You!

Questions?

- Our code is open source and publicly available at the github: https://github.com/hsqmlzno1/Transferable-E2E-ABSA

References

Niklas Jakob and Iryna Gurevych. 2010. Extracting opinion targets in a single- and cross-domain setting with conditional random fields. In EMNLP, pages 1035–1045.

Fangtao Li, Sinno Jialin Pan, Ou Jin, Qiang Yang, and Xiaoyan Zhu. 2012. Cross-domain co-extraction of sentiment and topic lexicons. In ACL, pages 410– 419.

Ying Ding, Jianfei Yu, and Jing Jiang. 2017. Recurrent neural networks with auxiliary labels for cross- domain opinion target extraction. In AAAI, pages 3436–3442.

Wenya Wang and Sinno Jialin Pan. 2018. Recursive neural structural correspondence network for cross- domain aspect and opinion co-extraction. In ACL, pages 2171–2181.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In KDD, pages 168–177.

Cigdem Toprak, Niklas Jakob, and Iryna Gurevych. 2010. Sentence and expression level annotation of opinions in user-generated discourse. In ACL, pages 575–584.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. Semeval-2015 task 12: Aspect based sentiment analysis. In SemEval, pages 486–495.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In SemEval, pages 27–35.

References

Maria Pontiki, Dimitris Galanis, Haris Papageor- giou, Ion Androutsopoulos, Suresh Manandhar, AL- Smadi Mohammad, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphe ́e De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In SemEval, pages 19–30.

Xin Li, Lidong Bing, Piji Li, and Wai Lam. 2019a. A unified model for opinion target extraction and target sentiment prediction. In AAAI, pages 6714–6721.

Joey Tianyi Zhou, Hao Zhang, Di Jin, Xi Peng, Yang Xiao, and Zhiguo Cao. 2019b. Roseq: Robust se- quence labeling. IEEE Transactions on Neural Net- works and Learning Systems, pages 1–11.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks? In NIPS, pages 3320–3328.

Lili Mou, Zhao Meng, Rui Yan, Ge Li, Yan Xu, Lu Zhang, and Zhi Jin. 2016. How transferable are neural networks in NLP applications? In EMNLP, pages 479–489.