Opinion-based Relational Pivoting for Cross-domain Aspect Term Extraction

Anonymous EMNLP submission

Abstract

Domain adaptation methods often exploit domain-transferable input features, a.k.a. pivots. The task of Aspect and Opinion Term Extraction presents a special challenge for domain transfer: while opinion terms largely transfer across domains, aspects change drastically from one domain to another (e.g. from *restaurants* to *laptops*). In this paper, we investigate and establish empirically a prior conjecture, which suggests that the linguistic relations connecting opinion terms to their aspects transfer well across domains and therefore can be leveraged for cross-domain aspect term extraction. We present several analyses supporting this conjecture, via experiments with four linguistic dependency formalisms to represent relation patterns. Following, we present an aspect term extraction method that drives models to consider opinion–aspect relations via explicit multitask objectives. This method provides significant performance gains, even on top of a prior state-of-the-art linguistically-informed model, which are shown in analysis to stem from the relational pivoting signal.

1 Introduction

Sentiment Analysis is one of the most widely used applications of natural language processing. A common fine grained formulation of the task, termed Aspect Based Sentiment Analysis, matches the terms in the text expressing *opinions* to corresponding *aspects*. For example, in the restaurant review in Figure 1, *great*, *calm* and *quiet* are opinion terms (OTs) referring to the aspect term (AT) *ambience*.

Following the SemEval shared tasks (Pontiki et al., 2014, 2015), the preliminary task of AT and OT extraction has attracted significant research attention (Wang and Pan, 2020; Pereg et al., 2020, inter alia), especially for its domain adaptation setup, where a model trained on one domain is tested on another, unseen domain. Domain adaptation is crucial for this task, since each product or service is considered a different domain. Yet performance on cross-domain aspect term extraction is still low, reflecting that this task poses a special challenge to common domain adaptation paradigms.

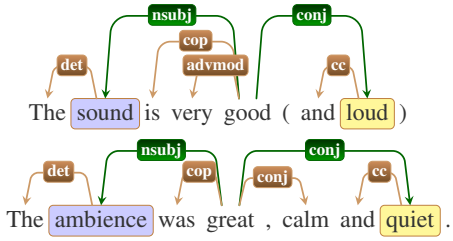

Figure 1: An OT (yellow) to AT (blue) path-pattern (green), defined on top of Universal Dependencies (UD), occurring in sentences from the *Devices* (top) and *Restaurants* (bottom) domains.

In most domain adaptation settings, some features of the input are domain specific, while others — also known as *pivot features* (Blitzer et al., 2006) — do transfer into unseen domains. Hence, cross-domain generalization concerns focusing the model's learning on the latter. However, aspect terms across domains share little direct commonalities. Essentially, the primary property aspects share is being the target topic referred to by opinion terms. For this reason, prior works suggested using hand-crafted syntactic rules (Hu and Liu, 2004; Ding et al., 2017), or alternatively, injecting a full syntactic analysis into the model (Wang and Pan, 2018; Pereg et al., 2020), aiming to capture the transferable relation-based properties of aspects.

Our first contribution is establishing the *relational pivoting* approach for cross-domain AT extraction on quantitative, data driven analysis (§3). We utilize four different linguistic formalisms (i.e., syntactic and semantic dependencies) to identify and characterize OT–AT relations, and empirically confirm their domain transferability and importance for the task. Following, we propose an auxiliary multi-task learning method with specialized relation-focused tasks, designed to teach the model

to focally capture these relations during OT and AT extraction training (§4). Our method improves cross-domain AT extraction performance over both vanilla BERT (Devlin et al., 2019) and the state-of-the-art SA-EXAL (Pereg et al., 2020) models. We conclude with a quantitative analysis of model predictions, ascribing observed performance gains to enhanced relational pivoting.

2 Background

Following the SemEval Aspect Based Sentiment Analysis shared tasks (Pontiki et al., 2014, 2015), recent works have formulated the *OT and AT extraction* task: given an opinionated text, identify the spans denoting OTs and ATs. We adopt the benchmark dataset that was used by recent works (Wang and Pan, 2020; Pereg et al., 2020), which consists of three customer-review domains — (R)estaurants, (L)aptops and digital (D)evices — and was aggregated from the SemEval tasks jointly with several published resources (Hu and Liu, 2004; Wang et al., 2016). While promising AT extraction performance has been demonstrated for in-domain settings (Li et al., 2018; Augustyniak et al., 2019), it does not scale to unseen domains, where state-of-the-art models exhibited small incremental improvements and struggle to surpass F1 scores of 40–55 (for the different domain pairs).

Previous works have conjectured that aspect and opinion terms maintain frequently occurring syntactic relations between them. Subsequently, Hu and Liu (2004), followed by Qiu et al. (2011), crafted a handful of simple syntactic patterns for in-domain AT extraction based on OTs. Motivated by the hypothesized domain transferability of syntactic OT–AT relations, Ding et al. (2017) employed pseudo labeling of AT based on the aforementioned patterns, which was used as auxiliary supervision for domain adaptation setup. We, however, exhaustively extract our patterns from the data rather than manually crafting them.

In a related line of work, syntax was leveraged more broadly for the same relational pivoting motivation. Wang and Pan (2018) and Wang and Pan (2020) encoded dependency relations with a recursive neural network using multitask learning, where the latter also applied domain-invariant adversarial learning. Most recently, the Syntactically Aware Extended Attention Layer model (SA-EXAL) (Pereg et al., 2020) improved cross-domain OT and AT extraction by augmenting BERT with an additional self-attention head that attends solely to the syntactic head of each token.

3 Motivating Analysis

The *Relational Pivoting* hypothesis is jointly entailed from two observations: (1) Opinion terms are similar across domains. (2) The relationships between corresponding OT–AT pairs have common, domain transferable linguistic characteristics. Taken together, these suggest that OT–AT linguistic relations are informative pivot features for transferring aspect extraction across domains. In the following subsections, we show several analyses supporting the above observations and hypothesis.

3.1 Opinions vs. Aspects Domain Variability

We first measure the degree to which OTs and ATs are shared across domains, by computing cross-domain lexical overlap. We find that while only 20% of aspect terms in target domain were seen at least once in source domain, this pertains to 69% of the opinions (see Appendix A for details). This is in sync with model experiments — both in-house and as reported by Wang and Pan (2020) — showing a drastic performance drop for cross-domain AT extraction, from lower 70s in domain to around 45 F1, while exhibiting a "reasonable" drop in OT extraction, from lower 80s to around 70 F1.

3.2 OT–AT Path Patterns

Next, we measure the degree to which linguistic relations connecting OT–AT pairs are shared across domains. To this end, we capture OT–AT linguistic relations using their *path pattern* in a dependency graph, i.e., the ordered list of the dependency relation labels occurring throughout the shortest (undirected) path between the terms (Figure 1).[1]

We investigate and compare four linguistic formalisms: **Spacy**'s syntactic dependencies[2], Universal Dependencies (**UD**), and two formalisms from Semantic Dependency Parsing (Oepen et al., 2015) — DELPH-IN MRS (**DM**) and Prague Semantic Dependencies (**PSD**).[3] We parsed all the sentences in

[1] We maintain edge direction by appending a directionality marker to each edge label. In case of multi-word terms, we take the token pair across the terms having the shortest path.

[2] https://spacy.io/

[3] We also experimented with three application-oriented UD extensions: Enhanced UD, Enhanced UD++ (Schuster and Manning, 2016), and pyBART (Tiktinsky et al., 2020). These formalisms introduced more label variability compared with UD, but also shortened OT–AT paths and performed slightly better in the multitask experiments. However, we omit these due to space considerations.

the benchmark dataset with state-of-the-art parsers — SpaCy 2.0, UDPipe[4], and HIT-SCIR (Che et al., 2019) for DM and PSD. Since correspondences between ATs and OTs are not annotated in the benchmark dataset, we consider for each formalism all OT–AT pairs whose shortest path length is ≤ 2. This yields 9K–10K pairs which cover 60%–70% of the ATs across the different formalisms. These pairs and their path patterns constitute the data for the analyses below, as well as for training relation-focused auxiliary tasks (§4).

We find that between 94%–97% of the patterns in one domain were seen at least once in another domain (more details in Appendix C). This confirms the prior presupposition that the linguistic structure of OT–AT relations is solidly domain invariant. Further, by plotting the cumulative distribution of patterns per domain, we observed that a small number of patterns covers a large proportion of pairs, especially for the semantic formalisms, which may facilitate learning of these patterns. Full plots and additional observations can be found in Appendix B. Altogether, these findings demonstrate that path-patterns are promising features for domain transfer.

3.3 Deterministic Relational Pivoting

To estimate the potential of relation-based pivoting, we analyze a deterministic approach for extracting ATs via gold OTs based on path patterns, similar to prior rule-based methods (Hu and Liu, 2004; Qiu et al., 2011), and assess how well such an approach transfer across domains. Given predicted linguistic parses, we select the top k common OT-to-AT path patterns and apply them on every OT, where traversal destination tokens are selected as ATs. To illustrate, given the UD pattern $\text{OT} \xleftarrow{\text{CONJ}} * \xrightarrow{\text{NSUBJ}} \text{AT}$, the OTs *quiet* and *calm* would both yield *ambience* as an AT (Figure 1, bottom). Notably, this analysis is only a rough upper-bound estimate; it is limited to identifying single-word ATs (70% of all ATs) which furthermore relates to an OT in a strictly known pattern, whereas models may generalize over some of these limitations.

Averaged results (across domain settings) are shown in Table 1 for varying k sizes.[5] Overall, pattern-based AT extraction can bring averaged F1 score up to **39** (DM), and recall up to **54** (UD). Crucially, there is hardly any drop in cross-domain settings relative to in-domain, affirming that patterns from a different source domain are as informative as in-domain patterns for opinion based AT extraction, consistent with observed pattern stability (§3.2). These findings suggest that driving a model to encode OT–AT relations should enhance domain adaptation.

	In-Domain						Cross-Domain					
	$k = 10$			All			$k = 10$			All		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
UD	41	32	35	22	54	31	40	30	34	22	52	30
DM	46	34	39	31	46	37	46	34	38	31	45	36

Table 1: Results of deterministically applying the top k common path patterns (in source domain) on gold OTs for extracting ATs. Evaluation is macro-averaged over the 3 in-domain or 6 cross-domain settings.

4 Multi-task Learning Method

To propagate the relational pivoting signal into an OT and AT extraction model, we apply auxiliary multitask learning (AMTL). We experimented with two auxiliary tasks for steering the model to encode OT–AT relationship information. Given an OT from an OT–AT pair of the collected auxiliary training data (§3.2), the model learns to: (1) predict its counterpart AT (ASP); and (2) predict the path-pattern connecting them on the dependency graph (PATT).[6] The ASP task should foreground the implicit representation of OT–AT relations, whereas PATT injects explicit, linguistically-oriented relation information.

Prior multitask learning approaches for enriching models with syntax (Strubell et al., 2018; Wang and Pan, 2018, 2020) have pushed them to encode a full syntactic analysis, possibly including irrelevant information. In contrast, our auxiliary tasks form a "partial parsing" objective, specialized in the relevant terms and their multifarious relations. We use both vanilla BERT (Devlin et al., 2019) and state-of-the-art SA-EXAL (Pereg et al., 2020) as base models, where the latter may imply whether our relation-focused signal is subsumed by SA-EXAL's awareness to the full syntactic parse (§2).

Implementation details We follow the experimental setup of (Pereg et al., 2020), including reported hyper-parameters, and apply one-layer classifiers on top of either `bert-base-uncased` or SA-EXAL encoders, both for OT and AT extraction (formulated as a BIO-tagging task), and for

[4] https://ufal.mff.cuni.cz/udpipe

[5] See Appendix D for analysis of domain pair variability.

[6] The SA-EXAL model was amended to generalize over the graph structures (rather than trees) produced by semantic formalisms (Appendix E).

	L → R	D → R	R → L	D → L	R → D	L → D	Mean
BERT	47.2 (4.0)	51.6 (2.1)	44.5 (3.1)	**46.7 (1.7)**	38.3 (2.4)	42.6 (0.6)	45.16
BERT + ASP — DM	**53.5 (3.3)**	**52.0 (2.1)**	45.7 (2.4)	45.9 (2.3)	38.8 (1.5)	**42.8 (1.0)**	**46.45**
BERT + ASP — Spacy	49.8 (3.2)	51.6 (1.5)	**46.2 (2.5)**	45.2 (2.5)	**39.4 (1.6)**	42.5 (1.0)	45.77
BERT + PATT — DM	46.3 (4.7)	50.9 (2.6)	42.9 (3.4)	46.2 (2.4)	38.0 (1.9)	42.1 (1.0)	44.40
BERT + PATT — Spacy	50.1 (3.0)	51.6 (2.0)	43.1 (2.2)	46.6 (2.5)	37.8 (1.6)	42.0 (0.9)	45.20
SA-EXAL — DM	48.7 (5.8)	53.8 (2.8)	46.0 (3.1)	**47.7 (1.8)**	**40.7 (1.3)**	41.9 (0.6)	46.48
SA-EXAL — Spacy	47.9 (3.1)	54.1 (1.9)	45.4 (3.3)	47.1 (1.1)	**40.7 (1.7)**	42.1 (1.4)	46.24
SA-EXAL + ASP — DM	**54.1 (2.3)**	51.6 (2.0)	45.6 (2.9)	45.8 (4.1)	39.2 (1.9)	41.8 (0.9)	46.37
SA-EXAL + ASP — Spacy	54.0 (3.1)	52.6 (1.9)	47.1 (3.0)	46.9 (2.4)	39.1 (2.7)	**42.2 (0.6)**	47.00
SA-EXAL + PATT — DM	52.8 (4.3)	**54.3 (1.8)**	**47.5 (1.9)**	**47.7 (2.2)**	40.3 (1.5)	41.6 (0.8)	**47.37**
SA-EXAL + PATT — Spacy	51.2 (3.4)	53.3 (2.3)	46.5 (2.3)	46.6 (1.8)	39.5 (1.2)	41.5 (0.9)	46.42

Table 2: Models evaluation (mean F1 score and standard deviation) for different datasets and linguistic formalisms. The best BERT and SA-EXAL results for each dataset are highlighted in bold.

the auxiliary tasks. Let $Z = \{z_1, z_2, \ldots, z_n\}$ be the contextualized representations of the input sequence produced by the encoder, and *op* be the OT index from an extracted OT–AT pair. The auxiliary classifiers are defined as follows:

$$\text{PATT}(Z, op) = softmax(z_{op}W^P + U^P)$$
$$\text{ASP}(Z, op) = softmax(o_1, \ldots, o_n)$$
$$o_i = (z_{op}W^A + U^A) \cdot z_i$$

where $W^P \in \mathbb{R}^{d \times m}$, $U^P \in \mathbb{R}^m$, $W^A \in \mathbb{R}^{d \times d}$, $U^A \in \mathbb{R}^d$ are model parameters, $\cdot$ stands for dot product, d is the hidden vector size and m is the size of the output pattern vocabulary. m is set by taking all the patterns whose frequency in training data (i.e., source domain) is ≥ 3, while mapping other patterns to a fixed UNK symbol.

5 Results and Analysis

Following (Pereg et al., 2020), we run each model on 3 random data splits and 3 different random seeds, presenting the mean F1 (and standard deviation) of the 9 runs. Detailed results are shown in Table 2, omitting the UD and PSD formalisms — which perform virtually on par with the other formalisms — for space considerations.[7]

For BERT, training for ASP consistently improves the mean F1 score, by up to 1.3 points (DM), bringing BERT's performance to be on par with the state-of-the-art SA-EXAL model. Improvements over the SA-EXAL baseline is generally smaller, yet some settings improve by 0.5–1 mean F1 points. Best performance is attained using SA-EXAL + PATT with semantic formalisms, indicating that pattern-focused signal is complementary to generic syntax enrichment methods.

[7]Results for models trained with both ASP and PATT were also omitted due to their lower performance.

Performance Analysis Examining the overlap between model predictions and the deterministic relational pivoting method (§3.3) can indicate to what extent the model utilizes relational pivot features. Specifically, we define *pivot-*ΔR as the recall improvement a model gains by unifying its true predicted ATs with those of the deterministic method (at $k = 10$).[8] Greater *pivot-*ΔR indicates greater discrepancy from the potential scope of pattern-based coverage. Taking **DM** as the formalism, we find that for the vanilla **BERT** model, average *pivot-*ΔR across 6 domain transfers is **16.5** recall points, with **22.6** for the Laptops to Restaurants transfer (L → R). This implies that relational features have a significant potential for enhancing its cross-domain coverage, especially on L → R, where we indeed observe the most profound model improvements using our relation-focused tasks. In comparison, **BERT + ASP** (DM) has an averaged *pivot-*ΔR of **14**, with **15.7** on L → R,[9] confirming that the multi-task objective pushes the model to cover more ATs using relational pivoting.

6 Conclusion

We establish an opinion-based cross-domain AT extraction approach, by analyzing the domain invariance of linguistic OT–AT path pattern. We consequently propose a relation-focused multitask learning method, and demonstrate that it enhances models results by utilizing relational features.

[8]We average this measure as well over the 9 runs of each model (#data splits × #seeds).

[9]More generally, the difference between *pivot-*ΔR for **BERT** and **BERT + ASP** (DM) correlates with model's improvement along the datasets (Spearman's $\rho = 0.83$).

References

Łukasz Augustyniak, Tomasz Kajdanowicz, and Przemysław Kazienko. 2019. Comprehensive analysis of aspect term extraction methods using various text embeddings. *arXiv preprint arXiv:1909.04917*.

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.

Wanxiang Che, Longxu Dou, Yang Xu, Yuxuan Wang, Yijia Liu, and Ting Liu. 2019. Hit-scir at mrp 2019: A unified pipeline for meaning representation parsing via efficient training and effective encoding. In *Proceedings of the Shared Task on Cross-Framework Meaning Representation Parsing at the 2019 Conference on Natural Language Learning*, pages 76–85.

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.

Ying Ding, Jianfei Yu, and Jing Jiang. 2017. Recurrent neural networks with auxiliary labels for cross-domain opinion target extraction. In *Association for the Advancement of Artificial Intelligence*, pages 3436—3442.

Minqing Hu and Bing Liu. 2004. Mining opinion features in customer reviews. In *American Association for Artificial Intelligence*.

Xin Li, Lidong Bing, Piji Li, Wai Lam, and Zhimou Yang. 2018. Aspect term extraction with history attention and selective transformation. *CoRR*, abs/1805.00760.

Stephan Oepen, Marco Kuhlmann, Yusuke Miyao, Daniel Zeman, Silvie Cinková, Dan Flickinger, Jan Hajic, and Zdenka Uresova. 2015. Semeval 2015 task 18: Broad-coverage semantic dependency parsing. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 915–926.

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, Dimitrios Galanis, 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) at (COLING 2014)*, pages 27–35.

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)*, pages 486–495.

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

Sebastian Schuster and Christopher D Manning. 2016. Enhanced english universal dependencies: An improved representation for natural language understanding tasks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2371–2378.

Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. 2018. Linguistically-informed self-attention for semantic role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5027–5038.

Aryeh Tiktinsky, Yoav Goldberg, and Reut Tsarfaty. 2020. pybart: Evidence-based syntactic transformations for ie. *arXiv preprint arXiv:2005.01306*.

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*, pages 1—-11.

Wenya Wang and Sinno Jialin Pan. 2020. Syntactically meaningful and transferable recursive neural networks for aspect and opinion extraction. *Computational Linguistics*, 45(4):705–736.

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.

A Cross–domain lexical overlap in Opinion and Aspect Terms

Table 3 shows the percentage of term instances in the target domain occurring at least once in the source domain. Overall, unlike aspect terms, opinion terms have significant overlap across domains. For example, the terms *great*, *good*, *best*, *better* and *nice* all occur in the top-10 common OTs in each of the three domains, jointly covering 22%, 20% and 14% of OTs in the *Restaurants*, *Devices* and *Laptops* domains, respectively. In sharp contrast, there is only one aspect (*price*) occurring in the top 50 common ATs at all three domains.

	Aspects	Opinions
D → R	7.3	78.6
D → L	42.3	83.2
R → D	12.2	59.1
R → L	11	61.4
L → D	41.3	65.4
L → R	9.1	68.3
Mean	**20.5**	**69.3**

Table 3: Cross-Domain lexical term overlap — how many term instances from target domain occur at least once in source domain (percentage).

B Pattern distribution analysis

As mentioned in Section 3.3, we plot the relative cumulative pattern distribution in each domain for each formalism, visualizing the number of different patterns versus OT–AT pairs coverage[%] (Appendix Figure 2). The general picture is that the vast majority of OT–AT pairs exhibit a few dozens of path patterns. Nevertheless, it is illuminating to examine the differences between domains and linguistic formalisms.

Specifically, we observe that the *Laptops* domain is the most diverse and slowly-accumulating, presumably making it harder to benefit from relational pivoting, while the opposite is true for the *Restaurants* domain. By and large, this is inline both with results of the deterministic pivoting analysis (Section 3.3) broken down by domain pairs (Table 4), and, to a smaller degree, with performance gains of our relation-focused multitask experiments (Section 5).

In addition, we find the cumulative distributions of **DM** and **PSD** to be more "dense". In DM, for example, the most frequent common pattern (simply OT$\xrightarrow{\text{ARG1}}$AT) covers 55% of the paths. This implies that semantic formalisms, designed for abstracting out surface realization details, strengthen the commonalities across different sentences, thus might have greater potential for relational pivoting. This conjecture is also backed by the deterministic pivoting analysis (Section 3.3). However, we did not find a significant advantage for semantic vs. syntactic formalisms using our multitask methods (See Section 5).

C Cross-domain overlap in path patterns

In Table 5, we present the percentage of target domain path patterns occurring at least once in the source domain. To account for pattern frequency in each domain, we also compute the Jensen-Shannon Distance between pattern probability distributions (Table 6). Overall, DM has the best cross-domain pattern overlap, while the *Devices* and *Laptops* domains are slightly more similar to each other.

D Deterministic relation pivoting per domain pair

In Section 3.3 we describe a deterministic DA AT extraction method based on gold opinion terms and top k most frequent OT–AT path patterns. Results per domain pair are shown in Table 4 for $k = 10$, which approximately optimizes recall-precision trade-off. Noticeably, the method is significantly less effective for the *Laptops* target domain. This finding aligns with its wider pattern diversity illustrated in Figure 2, but should also be attributed to it having relatively fewer OT–AT pairs that exceed our path-length ≤ 2 criterion. In **DM**, for example, the ratio of the number of selected OT–AT pairs to the total number of aspect terms is **0.93** for *Restaurants*, **0.77** for *Devices*, but only **0.67** for the *Laptops* domain. Altogether, our investigation suggests that the domains vary in linguistic complexity, reflected in richer and longer path patterns for truly corresponding OT–AT pairs in some domains (e.g. *Laptops*) compared to others (e.g. *Restaurants*). Relational pivoting might be more contributive to the latter, as also demonstrated by the multitask experiments (Section 5).

E SA-EXAL for semantic graphs

As mentioned in Section 2, the SA-EXAL model augments BERT with anspecialized,13th attention head, incorporating the syntactic parse directly into the model attention mechanism. In the original paper, SA-EXAL was fed with syntactic dependency trees, where each token has a syntactic head token to which it should attend. The learned attention matrix $A \in \mathbb{R}^{n \times n}$ is multiplied element-wise by a matrix representation of the syntactic parse P, where each row is a one-hot vector stating the token to which to attend.

However, semantic dependency formalisms, such as PSD and DM, produce bi-lexical directed acyclic graphs, in which a word can have zero "heads" (for semantically vacuous words, e.g. copular verbs) or multiple "heads" (i.e. outgoing edges). We modify the SA-EXAL model such that instead of one-hot rows, P can have all-one rows (no heads) or multiple-ones rows (multiple heads). Consequently, for tokens with no heads the network is learning the attention without external interfer-

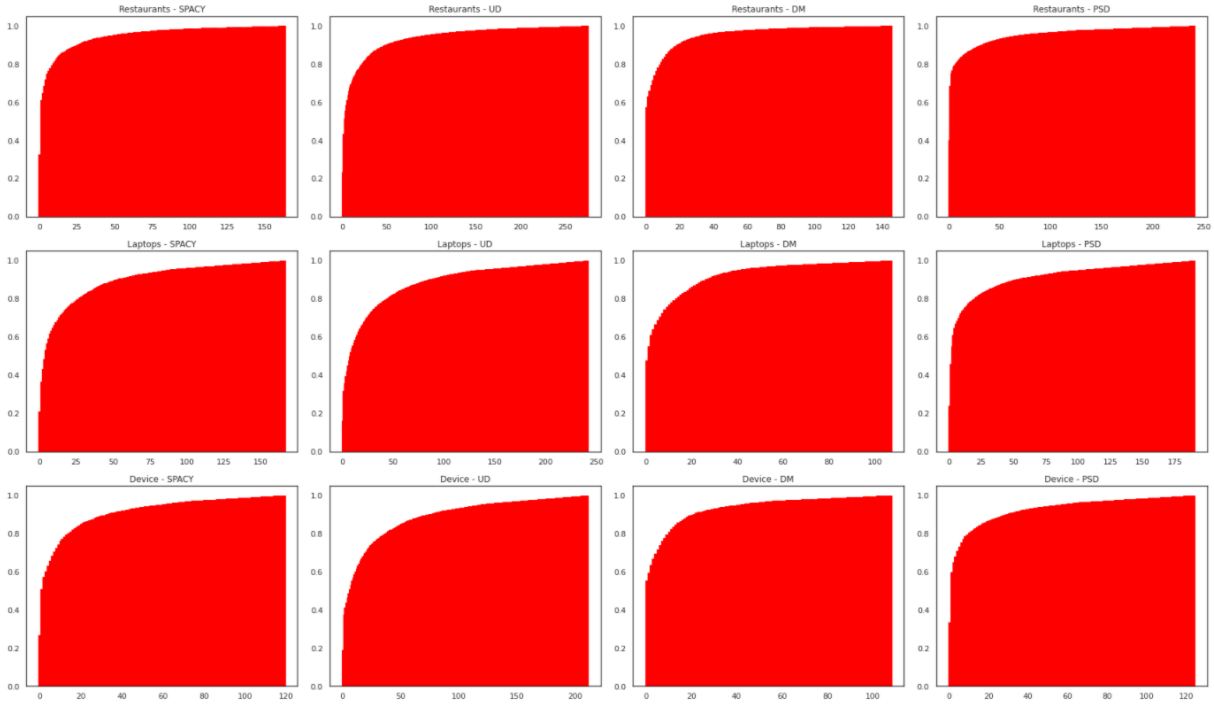

Figure 2: Relative cumulative frequency distributions of path patterns for each domain in all formalisms, showing how many different patterns (X axis) cover what percentage of OT–AT pairs (Y axis).

	R → L	R → D	L → R	L → D	D → R	D → L
spacy	P: 0.32 R: 0.22 F1: 0.26	P: 0.61 R: 0.29 F1: 0.4	P: 0.49 R: 0.37 F1: 0.42	P: 0.58 R: 0.33 F1: 0.42	P: 0.54 R: 0.37 F1: 0.44	P: 0.3 R: 0.24 F1: 0.27
ud	P: 0.26 R: 0.23 F1: 0.24	P: 0.46 R: 0.29 F1: 0.36	P: 0.44 R: 0.39 F1: 0.41	P: 0.47 R: 0.3 F1: 0.36	P: 0.49 R: 0.37 F1: 0.43	P: 0.26 R: 0.23 F1: 0.24
dm	P: 0.29 R: 0.25 F1: 0.27	P: 0.6 R: 0.34 F1: 0.44	P: 0.52 R: 0.4 F1: 0.45	P: 0.6 R: 0.37 F1: 0.46	P: 0.47 R: 0.39 F1: 0.43	P: 0.26 R: 0.26 F1: 0.26
psd	P: 0.22 R: 0.26 F1: 0.24	P: 0.41 R: 0.34 F1: 0.37	P: 0.35 R: 0.4 F1: 0.38	P: 0.41 R: 0.35 F1: 0.38	P: 0.3 R: 0.4 F1: 0.34	P: 0.19 R: 0.27 F1: 0.22

Table 4: Results of deterministic relational pivoting per DA settings (K=10).

	R → L	R → D	L → R	L → D	D → R	D → L
spacy	89.9	87.4	97.5	96.8	95.3	93
ud	93.4	94	96.7	95.9	95.7	93.1
dm	97.8	97.9	97.9	97	97.3	97.1
psd	93.8	95.5	95.3	96.8	93.2	90.4

Table 5: Cross-domain pattern overlap — how many AT–OT paths in target domain share a pattern with paths in source domain (percentage).

	R ↔ L	R ↔ D	L ↔ D	Mean
spacy	0.62	0.60	0.58	0.60
ud	0.60	0.59	0.56	0.58
dm	0.50	0.50	0.50	0.50
psd	0.60	0.56	0.58	0.58
Mean	0.58	0.56	0.55	0.57

Table 6: Jensen-Shannon Distances between pattern probabilities in different domains. Lower distance indicates similarity between the frequency signature of patterns in a domain pair.

ence, whereas for tokens with multiple heads, the attention mass is distributed between the heads.