Supplementary Material for EMNLP 2019 Paper: Leveraging Just a Few Keywords for Fine-Grained Aspect Detection Through Weakly Supervised Co-Training

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

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

A Datasets

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

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

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

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

The reviews of each language are already segmented into sentences. The average number of training and test segments across languages is around 2500 and 800 segments respectively. The training segments of restaurant reviews are significantly fewer than the training segments of product reviews. Therefore, for non-English reviews we report results after a single co-training round. For our co-training experiments we augment the English reviews dataset with 50,000 English reviews randomly sampled from the Yelp Challenge corpus.[1]

B Implementation Details

For a fair comparison, for the product reviews we use the 200-dimensional word2vec embeddings provided by Angelidis and Lapata (2018) and the base uncased BERT model.[2] For the restaurant reviews, we use the 300-dimensional multilingual word2vec embeddings provided by Ruder

[1] https://www.yelp.com/dataset/challenge
[2] https://github.com/google-research/bert#pre-trained-models

et al. (2016) and the multilingual cased BERT model.[3] The student's parameters are optimized using Adam (Kingma and Ba, 2014) with learning rate 0.005 and mini-batch size 50. After each co-training round we divide the learning rate by 10. We apply dropout in the word embeddings and the last hidden layers of the classifiers (Srivastava et al., 2014) with rate 0.5.

C More Results

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

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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