Teacher! Don't punish the Noisy Student.

Domain Generalization on ABSA task via utilizing noisy student architecture

Korea University COSE461 Final Project

Minjeong Ban*
Department of Industrial and
Management Engineering
Team 6
2020170805

Kyungho Kim*
Department of Industrial and
Management Engineering
Team 6
2020170815

Mina Hwang*
Department of Industrial and
Management Engineering
Team 6
2020170844

Hyewon Ryu*
Department of Industrial and
Management Engineering
Team 6
2020170845

Abstract

Recently, Large Language Model (LLM) have been used in many Natural Language Processing studies. LLM also utilized in ABSA task, which has been widely conducted based on BERT model. The InstructABSA, the background of our study, is a SOTA ABSA model based on T5 LLM model.

However, there is still limitation such as bias in specific domain. So, in this paper we aim to achieve domain generalization performance improvement on various domains by using our model called NoisyABSA(short for Noisy Student ABSA). It uses knowledge expansion concept by applying Teacher - Student model architecture using semi-supervised learning method.

The empirical results on two real-world datasets, SemEval 14 - Restaurant and Laptop, demonstrate the domain generalization performance of NoisyABSA, which outperforms the state-of-the-art (SOTA) model as shown in our experiment results.

1 Introduction

Sentiment analysis is a field that can be seen as the basic field of natural language processing. Sentiment analysis is the process of analyzing text to determine whether the emotional tone of the message is positive, negative, or neutral. Furthermore, Aspect-based sentiment analysis(ABSA) is the task that determines the polarity for each aspect such as a term or category in sentence , not for the whole sentence.

ABSA(short in Aspect-based sentiment anlysis) is usually applied to personal perspective data such as review data, survey response data, online data and social media data. It is important in field such as marketing and customer service.

The issues that affect our paper topic stem from the InstructABSA(Kevin S. et al. 2023) experiment results. InstructABSA is State-Of-The-Art (SOTA) model on all ABSA subtasks. It uses T5 model and instruction tuning for ABSA task. However, it shows poor performance on cross domain tasks, where data used in training and evaluation are from different domains. Performance of the model trained on cross domain task is even lower than the vanilla model(T5 model). It is a fatal limitation of InstructABSA because labeled data is cost expensive, and therefore is sparse in various domains.

Since linguistic characteristics that represent emotions are different between the various domains, it is important to improve generalization performance for ABSA. For example, two different domain reviews such as laptop review 'the battery performance is too high' and restaurant review 'the price is too high' have same term 'high' but represent different emotions because of different aspects.

In this paper, we aim to create a domain robust ABSA model by applying semi-supervised approach with Noisy Student(Qizhe Xie. et al. 2020) paper that improves performance on ImageNet classification. We will show training the InstructABSA with Noisy Student model architecture for enough iteration improves cross domain performance. We experiment on various situations for restaurant and laptop datasets with controlling two hyper-parameters, weight decay and max grad norm, as noise. Our model would be useful for domain generalization on ABSA tasks.

2 Related Work

2.1 InstructABSA

There have been many attempts to apply language models to the task of ABSA. The previous methods mainly utilized the BERT model, where the most recent approach was to train ABSA by Natural Language Inference(NLI) or Question Answering(QA) method via constructing auxiliary sentence (Chi S. et al. 2019)[1].

However, with the proposal of InstructGPT(Long O. et al. 2022), it has become a major issue in Natural Language Processing to utilize large language models with instructions. The paradigm has shifted from finetuning a language model to giving instructions and asking the large models. Accordingly, there are attempts to apply instruction tuning in ABSA. InstructABSA(Kevin S. et al. 2023) is the model that has achieved the SOTA performance for ABSA in 2023[2]. The InstructABSAis instruction tuned for three different subtasks of ABSA – ATE, ATSC, and Joint Task. The instruction prompt consists of the definition of the ABSA subtasks, along with two different positive, negative and neutral examples. Despite using a much smaller number of 200M parameters, the model outperformed the previous SOTA model. Thus, the study includes the cross and joint domain evaluation, leaving room for domain generalization as we mentioned in Introduction.

2.2 Noisy Student

We want to leverage the structure proposed in the Self-training with Noisy Student impressions ImageNet classification (Qizhe Xie. et al. 2020) paper, which was accepted in CVPR 2020 and achieved SOTA performance on the Image Classification Task at the time[3]. We improved our model by applying this model to our paper. The model is meaningful in making use of Semi-supervised learning to take advantage of the large number of unlabeled data that exists in the world. A brief description of the model's training process is as follows.

(1) Learn the Teacher Model with Labeled Data.

(2) To create a pseudo label of Unlabeled Data, using the Teacher model trained in (1).

(3) Combine Labeled Data and Unlabeled Data to train Student Models that are larger than or equal to the Teacher Model.

Student model trained in (3) recreate the Pseudo Label and act as the Teacher model, and train other students with the corresponding Pseudo label data. Repeating this process is the methodology proposed in the Noisy Student paper, and in the process of training Student, noise such as 'dropout', 'stochastic depth', and 'data augmentation' is added to train a more robust model. The model is consistent with the direction we want to proceed as we aims to improve robustness as well as accuracy. Therefore, we would like to propose a more robust model structure for each domain through the application of the model to ABSA Task.

3 Approach

3.1 Baseline+Prompt

In ABSA tasks, models should understand both implicit and explicit meanings, so it can be expected that models that have strength in NLI tasks will show high performance in ABSA tasks as well. Since large language models are showing strong performance on NLI task, we used the Tk-instruct, which is T5 based model, as the baseline of our model. Tk-instruct is a generative model which is based on T5 and trained by various in-context instructions[4]. This is exactly the same baseline of InstructABSA, which achieved SOTA before this paper.

In addition to this baseline, we fine tuned on downstream tasks of ABSA. We tried instruction tuning to increase the domain generalization capability. Our instruction prompts consist of task definition and two examples of positive, negative and neutral as our model focuses on ABSA tasks. Each example consists of examples for two different domains, restaurant and laptop . Detailed instructions are presented in Figure 1.

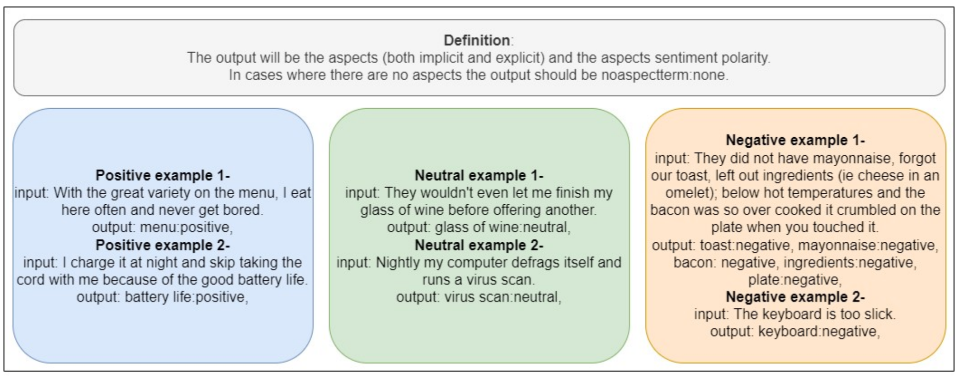


Figure 1: Prompt Example

3.2 Noisy Student

We propose our model pipeline by using Noisy Model's architecture mentioned at Related Work. The basic structure of the model consists of a structure that transfers the knowledge of the Teacher Model to the Student Model.

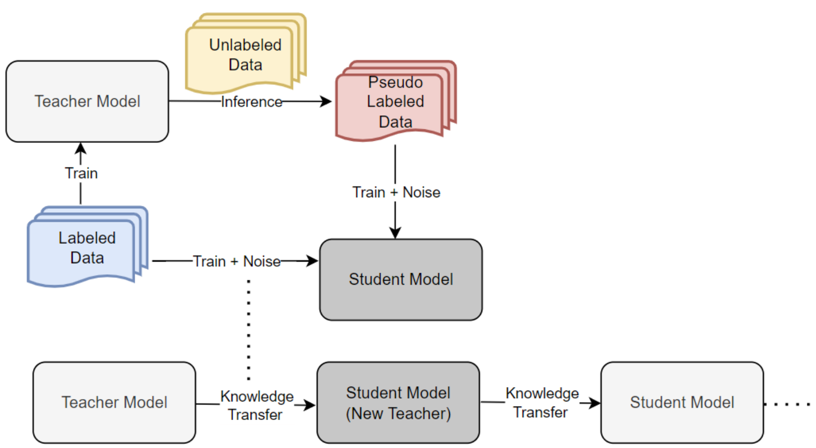


Figure 2: The Model pipeline of NoisyABSA

First, the Teacher Model is trained on the Labeled Dataset. Then unlabeled data is inferenced by the Teacher model learned through the previous step. The labels inferenced by the teacher model are called Pseudo Label. Using existing labeled data and pseudo labeled data, we train the Student Model that is the same in model size as the teacher model. In the process of training the Student Model, we would like to implement a more robust model by giving NOISE. The biggest reason for noise is that the pseudo labeled data inferred by the teacher model cannot be fully trusted. Therefore, we give noise through the adjustment of Hyper-parameters of Weight Decay and Max Grad Norm among various hyper-parameters in the model training process.

In the second iteration, the learned student model becomes the new Teacher and performs a new inference to unlabeled data, and trains a new student model using the newly inferred pseudo-labeled data and labeled data. At this time, noise is also given. By iterating this process, the final, well-learned student model is created. The process is described in Figure 2.

4 Experiments

4.1 Data

SemEval 2014 Task 4 dataset is used for our experiment. The dataset is used as a benchmark for ABSA tasks and has customer reviews from two domains - laptops and restaurants. The dataset contains train, test, trial dataset for each domain. The train and trial sets are labeled data, consisting of raw text and the aspect term, polarity pair. The test set is unlabeled data, consisting only of raw text. We obtained the dataset from Kaggle.

Input Dataset Example	
raw_text	Everything is so easy to use, Mac software is just so much simpler than Microsoft software.
aspectTerms	[{'term': 'Mac software', 'polarity': 'positive'}, {'term': 'Microsoft software', 'polarity': 'negative'}, {'term': 'use', 'polarity': 'positive'}]

Figure 3: Example of Input Dataset

We adjusted the original dataset to construct new sets of train, unlabeled, and test dataset for our NoisyABSA. The train and trial dataset from the original dataset were used to create our train and test dataset. We took 300 samples to construct the test dataset, and used the rest for training. The original test dataset is used as the unlabeled dataset for our study. The detailed dataset statistics are presented in Table 1.

Train data is used for training the teacher model. Unlabeled data is fed to the teacher model for inference, and the teacher model labels the unlabeled data to create pseudo labeled data, which is then used for training the student model. Test data is used to evaluate the performance of the teacher and student models.

	Restaurant	Laptop
Train	1773	1222
Unlabeled	800	800
Test	300	300

Table 1: Dataset Statistics

4.2 Experimental details

4.2.1 Domain Tasks

We conduct our experiment on three novel domain tasks: Cross, Semi-Cross1 and Semi-Cross2. We create three different domain tasks according to various domain combination of train, unlabeled and test dataset.

Cross Domain Task refers to when the domain of the test dataset is different from the train, and unlabeled dataset. For example, the training and unlabeled datasets are reviews of restaurants, whereas the test dataset consists of reviews of laptops. Compared to the cross domain proposed in the InstructABSA paper, the train dataset is expanded by applying noisy student structure with unlabeled datasets.

To achieve better domain generalization performance, we adopted two new domain tasks, Semi-Cross1 and 2. In Semi-Cross1 domain task, the teacher model is trained with train data on domain A and unlabeled data on domain B, then tested on domain B. We believe that using unlabeled data from another domain B would give better domain generalization performance than to train only with domain A.

Semi-Cross2 domain task is the same as Semi-Cross1 except that unlabeled data on domain A is added to train the student model. We expect Semi-Cross 2 to show better performance than Semi-Cross1. The three domain tasks are stated more clearly in Table 2.

In addition to the three domains, we have conducted the experiment on joint domain where both domains of laptop and restaurant were used for train and unlabeled set, then tested on one of those domains. As expected, the teacher and student model (on iteration 1) for the joint domain showed a good performance. However, because using both domains for training is against the purpose of our research goal, domain generalization, we excluded joint domain from our overall experiments.

Domain Task	Train	Unlabeled	Test
Cross	A	A	B
Semi-Cross1	A	B	B
Semi-Cross2	A	A, B	B
Joint	A, B	A, B	B

Table 2: The domain of the train, unlabeled, and test datasets differ according to the domain category. A and B are two different domains - either laptop and restaurant each, or restaurant and laptop.

4.2.2 Parameters

In the paper of Noisy Student on Image Classification task, they added noise by data augmentation(data noise), dropout, stochastic depth(model noise). But in Natural Language Processing task especially ABSA task, noise by data augmentation cannot be used. Here is the reason. One of the biggest features of data augmentation is that the way it works depends on the domain of the data. In the case of computer vision, most of them deal with two dimensional images, and when solving classification problems, they take advantage of the fact that the label does not change easily when the image is cut of transformed. However, in the case of NLP, the various methodologies of computer vision cannot be applied immediately because changing a word can completely chage the meaning of the sentence. For example, in the sentence “This is the best movie I’ve ever seen”, if the word ‘best’ changed to ‘worst’ the meaning of the sentence changes completely.

Therefore, we give just model noise by controlling weight decay or max grad norm parameter rather than dropout because there is no dropout hyper-parameter in training model.

For our noisy student model, we wanted to give some noise to the student model by letting it lose some information. Therefore, we tested on two parameters: weight decay and max grad norm. The default setting for weight decay is 0.01 and for max grad norm 1.0. We chose to employ weight decay = 0.1 and max grad norm 3.0 to experiments.

4.2.3 Iterations

For our overall teacher student structure, we wanted to make the student model even smarter and smarter by repeating the process of knowledge transfer. Therefore we conducted the experiment for 1 to 4 iterations, to find out the optimal number of iterations.

4.3 Evaluation Metric

The quantitative evaluation metric used in the experiment is F1 score. However, when checked qualitatively, it was found that some of the correct answers were regarded as incorrect due to the mismatch of adjectives in the aspects.

Therefore, we employ an additional process to correct the accuracy of the prediction labels as stated in the InstructABSA paper. When the predicted term and the ground truth term were in an inclusion relationship, we treated them as correct answer.

Text	Ground Truth	Predicted
Toshiba is aware of the issue but unless the extended warrenty is bought Toshiba will do nothing about it.	extended warranty	warranty
There is no number pad to the right of the keyboard which is a bummer.	pad	number pad
Now I had not tried to use this since the disc drive had been replaced and after taking it back to the Geek Squad I found out they had accidently not used the right drive when they replaced the first one, so back it went to get the correct drive.	drive	disc drive

Table 3: Examples of the aspect terms pair which is consisted with 'ground truth term' and 'predicted term'.

F1 score	Restaurant		Laptop	
	Teacher	Student	Teacher	Student
Cross	0.6371	0.6500	0.5681	0.5771
Semi-Cross1	0.6371	0.6484	0.5681	0.5726
Semi-Cross2	0.6371	0.6472	0.5681	0.5869

Table 4: F1 score of one iteration and default parameters for teacher and student model in each domain tasks.

4.4 Results

4.4.1 Teacher Student Structure

Table 4 denotes the result of Cross, Semi-Cross1 and 2 domain task for restaurant and laptop datasets. The results are F1 score of one iteration and default parameters for teacher and student model.

Performance of student model is better than teacher model for all tasks and datasets. Thus, training the student model with inferred unlabeled data by teacher model improve the performance. This means NoisyABSA is reasonable.

4.4.2 Noisy Student

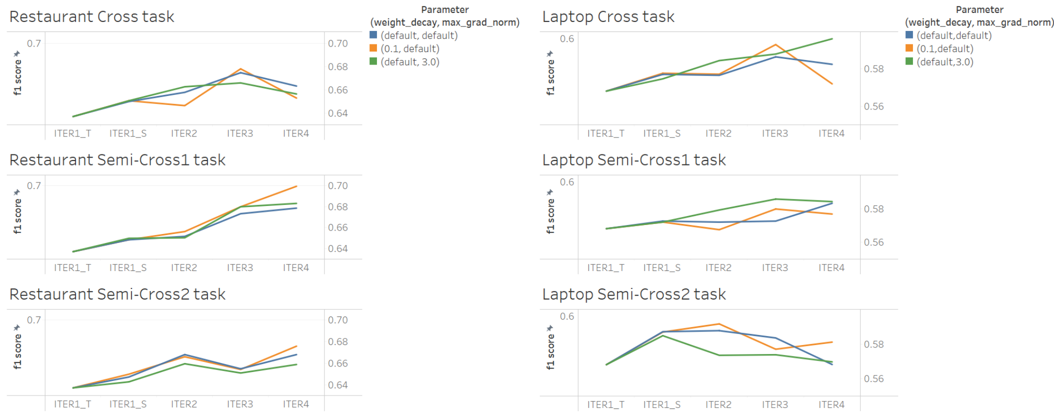
Figure 4 (a) denotes the result of Cross, Semi-Cross1 and 2 domain task for 3 different parameter cases and 4 iterations in restaurant dataset. Figure 4 (b) is the same same task in laptop dataset. The x-axis is the number of iterations and y-axis is F1 score. And colors of graphs mean different parameter combinations.

First, we tuned two different parameters to student be noisy. As a result, tuning parameters improves the performance. In 4 out of 6 tasks, the best performance appears when weight decay is 0.1 and max grad norm is default. And in 2 out of 6 tasks, model with 3.0 max grad norm results the best performance. Thus, being noisy can make better model.

Second, we also experiment teacher-student structure for 4 iterations. Second iteration means the first student become a new second teacher model and teach new second student. In other words, the pupil has become the master as the iteration continues. Usually the performance of $n(n \geq 2)$ iterations is greater than the performance of 1 iteration. This result means iteration of teacher-student structure is meaningful.

4.4.3 Comparison with InstructABSA

For all datasets in semeval-2014, our model NoisyABSA surpasses the existing SOTA model, InstructABSA, and achieves new SOTA in the cross domain task. Results are presented in Table 5. The score of InstructABSA is the result of re-experimentation on the datasets used in this paper through the InstructABSA implementation codes published by authors. Since NoisyABSA varies depending on the type of noise and the number iterations, the scores for NoisyABSA in Table 5 is the highest score achieved for each task.



(a) in restaurant domains

(b) in laptop domains

Figure 4: Performance comparison graphs by hyperparameter(weight decay, max grad norm) for three (Cross, Semi-Cross1, Semi-Cross2) tasks

Model	**Noise**		**Score**	
	Weight Decay	Max Grad Norm	Restaurant	Laptop
InstructABSA(Cross)	Default (0.01)	Default (1.0)	0.6371	0.5681
NoisyABSA(Cross)	0.1	Default	0.6781	0.5931
NoisyABSA(Cross)	Default	3.0	0.6660	**0.5962**
NoisyABSA(Semi-Cross1)	0.1	Default	**0.6993**	0.5798
NoisyABSA(Semi-Cross1)	Default	3.0	0.6830	0.5857
NoisyABSA(Semi-Cross2)	0.1	Default	0.6758	0.5916
NoisyABSA(Semi-Cross2)	Default	3.0	0.6595	0.5738

Table 5: F1 score table for each hyper-parameter. Presenting the best score of each model ignoring iterations.

In this way, NoisyABSA's improvement in cross domain task performance over existing SOTA model means that NoisyABSA can mitigate domain generalization performance degradation that can occur when large language models are trained for a specific downstream task.

5 Analysis

Some experiment results are slightly different from our expectations. We expected performance on semi-cross domain task will outperform cross domain task. In Restraunt domain, our hypothesis is satisfied but, in Laptop domain, cross domain task's performance outperform slightly than semi-cross domain's task. We discussed about the reason and the conclusion was as follows.

In the case of Laptop semi-cross domain contains a lot of more specific terms than Restaurant domain. After training Restaurant data, the accuracy of the inferred data of the unlabeled data may be poor. We concluded that the pseudo label was incorrectly inferred and it comes as a result of its poor performance than cross domain task.

We implemented InstructABSA in the same way with the paper, but the performance was shown as much lower. This is because the authors of the paper corrected the case where it was well predicted but evaluated as wrong due to a small difference in terms by performing a human-check on the output of the model. On the other hand, in this paper, only when the predicted term and the ground truth term were in an inclusion relationship was treated as correct answer.

6 Conclusion

In this work, we study the task of Aspect Based Sentiment Analysis. We propose a Noisy Student based solution, NoisyABSA that use Teacher - Student Architecture by transfer knowledge using semi-supervised learning method. The model's domain generalization performance outperform existing SOTA model, and our method seems to be robust to various domain through adding NOISE to student model.

As future work, we plan to get optimal common hyper-parameter(number of iteration, weight decay, etc) in noise to get more generalization framework in ABSA. In addition, we may use another instruction or large language model.

The significance of our model is as follows. (1) Our model showed the potential for Domain Generalization in ABSA Task. (2) Our model outperformed the existing SOTA model with our measured Evaluation Metrics. (3) We proposed a model structure that can utilize unlabeled data. Our model can be commonly used for various tasks in addition to ABSA tasks. Therefore, we expect to further extend our architecture to various tasks to mitigate the domain generalization performance reduction which occurs during training large language models to certain downstream tasks.

6.1 Limitation and Future Work

Previously, good performance appeared when weight decay was changed higher as noise and when the number of iterations was 2 or higher. However, in detail, the optimal parameters varied depending on the domain and task. Therefore, future studies should carry out more generalized common parameters through the search for more diverse hyper-parameters related to noise and sufficient number of iterations.

Also, in our work, we only focused on cross domain tasks. But, in order to accurately check performance of the model, it is necessary to measure the performances for individual domain tasks and joint domain tasks in addition to the cross domain task.

Furthermore, since the number of unlabeled dataset is larger than labeled dataset in Semi-Cross2 domain task, error of inferred unlabeled data by teacher model would be increase. As a result, the error also transfer to student model and it leads to low performance. Thus, we can expect better performance with enough train dataset.

Lastly, ABSA is a sensitive and complicated task because it has to extract several terms and its sentiment. NoisyABSA structure can be applied to another tasks and it is expected higher performance in simple tasks such as classification.

References

[1] Chi Sun, Luyao Huang, and Xipeng Qiu. Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence, 2019.

[2] Kevin Scaria, Himanshu Gupta, Siddharth Goyal, Saurabh Arjun Sawant, Swaroop Mishra, and Chitta Baral. Instructabsa: Instruction learning for aspect based sentiment analysis, 2023.

[3] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification, 2020.

[4] Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Maitreya Patel, Kuntal Kumar Pal, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Shailaja Keyur Sampat, Savan Doshi, Siddhartha Mishra, Sujan Reddy, Sumanta Patro, Tanay Dixit, Xudong Shen, Chitta Baral, Yejin Choi, Noah A. Smith, Hannaneh Hajishirzi, and Daniel Khashabi. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks, 2022.

A Appendix: Team contributions

We conducted entire project through collaborations and many discussions. Thus, all of our team members contributed to every part of project together. The parts that each person mainly played are as follows.

Minjeong Ban: She studied about BERT based ABSA and designed new domain tasks and experiment various combination of parameters.

Kyungho Kim: He studied model architecture of noisy student and InstructABSA and modified some codes in InstructABSA and implement experiments.

Mina Hwang: She studied about ABSA tasks and prompt design and experiment various combination of parameters.

Hyewon Ryu: She studied model architecture of InstructABSA and implemented codes for our model and designed experiments.

TASK	Weight decay	Max_grad_norm	TEST DOMAIN	ITER1_T	ITER1_S	ITER2	ITER3	ITER4
Cross	Default	Default	RESTAURANT	0.6028 (0.6371)	0.6142 (0.6500)	0.6155 (0.6579)	**0.6363 (0.6748)**	0.6208 (0.6633)
Cross	Default	Default	RESTAURANT	0.6028 (0.6371)	0.6142 (0.6500)	0.6155 (0.6579)	**0.6363 (0.6748)**	0.6208 (0.6633)
Cross	0.1	Default	RESTAURANT	0.6028 (0.6371)	0.6148 (0.6507)	0.6040 (0.6465)	**0.6337 (0.6781)**	0.6140 (0.6530)
Cross	Default	3.0	RESTAURANT	0.6028 (0.6371)	0.6148 (0.6507)	0.6225 (0.6627)	**0.6278 (0.6660)**	0.6159 (0.6565)
Cross	Default	Default	LAPTOP	0.5273 (0.5681)	0.5352 (0.5771)	0.5407 (0.5766)	0.5506 (0.5865)	**0.5531 (0.5825)**
Cross	0.1	Default	LAPTOP	0.5273 (0.5681)	0.5357 (0.5777)	0.5412 (0.5772)	**0.5589 (0.5931)**	0.5460 (0.5720)
Cross	Default	3.0	LAPTOP	0.5273 (0.5681)	0.5329 (0.5747)	0.5484 (0.5845)	0.5601 (0.5580)	**0.5645 (0.5962)**
Semi-cross1	Default	Default	RESTAURANT	0.6028 (0.6371)	0.6072 (0.6484)	0.6161 (0.6516)	0.6379 (0.6732)	**0.6509 (0.6785)**
Semi-cross1	0.1	Default	RESTAURANT	0.6028 (0.6371)	0.6072 (0.6484)	0.6208 (0.6562)	0.6405 (0.6798)	**0.6582 (0.6993)**
Semi-cross1	Default	3.0	RESTAURANT	0.6028 (0.6371)	0.6068 (0.6498)	0.6110 (0.6503)	0.6383 (0.6798)	**0.6497 (0.6830)**
Semi-cross1	Default	Default	LAPTOP	0.5273 (0.5681)	0.5394 (0.5726)	0.5428 (0.5720)	0.5396 (0.5726)	**0.5536 (0.5832)**
Semi-cross1	0.1	Default	LAPTOP	0.5273 (0.5681)	0.5389 (0.5720)	0.5382 (0.5675)	0.5464 (0.5798)	**0.5495 (0.5768)**
Semi-cross1	Default	3.0	LAPTOP	0.5273 (0.5681)	0.5389 (0.5720)	0.5458 (0.5792)	0.5524 (0.5857)	**0.5530 (0.5842)**
Semi-cross2	Default	Default	RESTAURANT	0.6028 (0.6371)	0.6044 (0.6472)	0.6229 (0.6680)	0.6233 (0.6548)	**0.6325 (0.6680)**
Semi-cross2	0.1	Default	RESTAURANT	0.6028 (0.6371)	0.6051 (0.6498)	0.6288 (0.6660)	0.6207 (0.6542)	**0.6426 (0.6758)**
Semi-cross2	Default	3.0	RESTAURANT	0.6028 (0.6371)	0.6000 (0.6427)	0.6125 (0.6595)	0.6135 (0.6509)	**0.6172 (0.6588)**
Semi-cross2	Default	Default	LAPTOP	0.5273 (0.5681)	0.5498 (0.5871)	0.5545 (0.5877)	**0.5546 (0.5835)**	0.5415 (0.5682)
Semi-cross2	0.1	Default	LAPTOP	0.5273 (0.5681)	0.5494 (0.5869)	**0.5578 (0.5916)**	0.5503 (0.5770)	0.5503 (0.5811)
Semi-cross2	Default	3.0	LAPTOP	0.5273 (0.5681)	0.5473 (0.5848)	0.5445 (0.5735)	**0.5490 (0.5738)**	0.5410 (0.5697)

Table 6: F1 score table for all experiments we conducted. The value written above is the score when considered correct only if the predicated term completely matches the ground truth term, and the value written in the lower bracket is the score when considered correct even if it is in an inclusion relationship.