Annotating Opinions and Opinion Targets in Student Course Feedback

Janaka Chathuranga, Shanika Ediriweera,
Ravindu Hasantha, Pranidhith Munasinghe and Surangika Ranathunga
Department of Computer Science and Engineering
University of Moratuwa, Katubedda 10400, Sri Lanka
{janaka.13, shanika.13, ravindu.13, pranidhith.13, surangika}@cse.mrt.ac.lk

Abstract

In this paper, we present a student course feedback corpus, a novel resource for opinion target extraction and sentiment analysis. The corpus is developed with the main aim of summarizing general feedback given by students on undergraduate-level courses. In this corpus, opinion targets, opinion expressions, and polarities of the opinion expressions towards the opinion targets are annotated. Opinion targets are basically the important key points in feedback that the students have shown their sentiment towards, such as "Lecture Slides", and "Teaching Style". The uniqueness of the corpus, annotation methodology, difficulties faced during annotating, and possible usages of the corpus are discussed in this paper.

Keywords: Corpus Annotation, General Student Feedback, Opinion Targets, Summarization, Sentiment

1. Introduction and Motivation

Student feedback is widely used in present in order to enhance the quality of teaching and learning process. Feedback is collected from students as online forms, as well as handwritten documents. Since it takes a considerable effort to read and understand all the feedback given by the students, the best way is to read all the feedback and create a summary that covers all the aspects of the given feedback.

Our intention was to design a system to automatically summarize student feedback for different aspects of a course, such as teaching style of the lecturer, learning environment, and presentation slides. The summarization process contains 3 phases.

1. Identifying and extracting all the opinion targets in the given feedback.
2. Clustering the opinion targets into unique categories.
3. Determining the sentimental polarity of the targets and getting a statistic of polarity for each target cluster.

In order to extract opinions and targets, the most promising technique is supervised learning (Luo & Litman, 2015; Luo, Liu, & Litman, 2016; Luo, Liu, W., Liu, F., & Litman, 2016). In order to apply the supervised machine learning techniques, we need a student course feedback corpus annotated with a suitable annotation scheme.

Our data corpus consists of student responses collected from an undergraduate Computer Science and Engineering course. General responses were collected from 20 lectures and workshops. They contain 973 student responses in total with 2,395 sentences.

In the annotation scheme, we annotate the opinion target and the opinion expression that shows the polarity of the target. For example, in the sentence "The lecture is really good", we annotate "lecture" as the opinion target and "really good" as the opinion expression. Polarity of the target "lecture" becomes positive by the annotated opinion expression.

We used the BIO (Beginning-Inside-Outside) (Sang & Veenstra, 1999) sequence labelling scheme to tag the annotated word phrases. BIO tagging includes "B" tags for the first word in a phrase, "I" tags for the other annotated words inside the phrase, and "O" tags for the words that are not annotated. We used BIO tagging because it is the most promising sequence labelling scheme that is used for many supervised machine learning models such as CRF (Conditional Random Fields) (Luo & Litman, 2015; Luo, Liu, & Litman, 2016; Luo, Liu, W., Liu, F., et al., 2016), in order to extract opinion targets and opinions.

This data corpus is unique because there are no such annotated corpora available for general student feedback. General Feedback means that feedback is taken for every aspect of the lecture by asking a question such as "Give feedback on today's lecture". Therefore, this data corpus contains both positive and negative opinions towards a target.

The rest of the paper is organized as follows. Section 2 describes about the previous research done on the similar contexts. The next section describes about the statistics of dataset and the data collection process. Section 4 describes the pre-processing steps that we carried out. Then Section 5 elaborates on the annotation scheme that we developed in order to feed the data to the machine learning models. Finally, Section 6 concludes the paper with an outlook into the future.

2. Related Work

In order to perform both aspect extraction and sentiment analysis using supervised learning techniques, the dataset has to be properly annotated. For annotation, either sequence labeling or sentence labeling based on aspects have been commonly used.

2.1 Sequence Labeling

Sequence labelling (Erdogan, 2010) is a type of pattern recognition task that involves the algorithmic assignment of a categorical label to each member of a sequence of observed values. In text analysis, tokens are taken as members.

Among the sequence labeling approaches, one of the most promising techniques to extract aspects/ opinion targets is Conditional Random Fields (CRF) (Lafferty, McCallum, & Pereira, 2001). In order to train a CRF model, word phrases are required to be annotated using BIO tags. Previous research (Turian, Ratinov, Bengio, & Turian, 2010) has used this labelling scheme for text chunking and Named Entity Recognition (NER) as well.

Other than BIO, POS tags can be identified as another labelling scheme. It has been used as a feature for highest scoring systems in SemEval-2015 Task 12: Aspect Based Sentiment Analysis (Zhang, Z., & Lan, M., 2015).

2.2 Sentence-level Labeling

In this approach, the dataset is evaluated sentence by sentence, categorizing them based on the aspect and the opinion into predefined labels or topics.

Supervised aspect extraction tasks have also used sentence-level labelling (Pontiki et al., 2016). Unlike sequence labelling that annotates the dataset inline, data used in sentence-level labelling (Pontiki et al., 2016) contain sentences each annotated with one of the pre-defined aspect categories. In instances where more than one aspect is found in a single sentence, the sentence is tagged with all the relevant aspect labels[1].

2. 3 Student Course Feedback

Student course feedback related previous work (Luo & Litman, 2015; Luo, Liu, & Litman, 2016; Luo, Liu, F., Liu, Z., et al.2016) has been done using reflective prompts. Since the polarity of the opinions is implied in the prompt itself, only opinion targets are annotated. Here, a sequence labeling scheme has been used. Moreover, reflective prompts have been designed carefully in such a way that only topics discussed in the class emerge as opinion targets. Human annotators have been used to divide feedback into relevant topics (Luo et al., 2016). In later work (Luo, Liu, & Litman, 2016), the annotation scheme has been improved by introducing a highlighting scheme that assigns a specific colour to similar topics. Then extractive methods such as Integer Linear Programming (Luo et al., 2016), phrase-based approach, clustering, and ranking approaches (Luo, Liu, & Litman, 2016) have been used to summarize student feedback. A dataset that has more than 900 responses for each reflective prompt has been used in this research.

In another work related to student feedback, Welch et al., (2016) have used a dataset consisting of 1,042 responses acquired from a Facebook student group. Only course names and instructor names were annotated as opinion targets with the respective polarity of the opinions towards them as positive or negative.

3. Data Collection

Our student course feedback corpus consists of student responses collected from an undergraduate Computer Science and Engineering Course in a South Asian university. Feedback given for lectures and workshops was collected through an online Learning Management System. The responses are anonymous and the language used is English. All the lectures were done by the same lecturer whereas each workshop was done by different presenters. Statistics of the student course feedback corpus are given in Table 1.

No. of Lectures	8
No. of Workshops	12
Total no. of student responses	973
Total no. of sentences	2,395
Avg. responses per lecture/workshop	48.65
Max. responses in a lecture/workshop	81
Min. responses in a lecture/workshop	7

Table 1: Statistics of the corpus

The prompts we used to collect responses were general prompts such as "What is your opinion about today's lecture?". Therefore, students had the freedom to write regarding any aspect of the lecture, unlike what they would write in response to a reflective prompt. In addition, there was no sentence limitation for providing feedback. The outstanding quality of this corpus is that it has student feedback that consists of both positive and negative opinions and comments about all the aspects of the course such as lecturer's qualities, course material, course content, and learning environment.

The sentiment of feedback about certain aspects such as "lecturer" tends to be biased towards positive. This nature could be a characteristic of south Asian cultural background.

4. Pre-Processing

In a country where English is considered as a 3rd language, it is not so surprising for students to make many errors when writing in English. Furthermore, unlike in other situations (e.g. writing a project report), since while giving feedback students deliberately do not worry about using correct English, there were many spelling mistakes in the corpus. Out of the 973 responses, 44.3% of the responses had spellings mistakes. In certain cases, we noticed that the meaning of the complete sentence gets distorted because of the spelling mistakes. For example, some students have used "Work shop is fine" instead of "Workshop is fine", which gives a very different meaning compared to what is intended by the student.

[1]SemEval-2016 Task 5 Available at: http://alt.qcri.org/semeval2016/task5/index.php?id=data-and-tools

Therefore, correcting the spelling mistakes beforehand the annotation phase became very critical for target and sentiment extraction. Thus, we carried out context-sensitive spell correction ("Microsoft Cognitive Services— Bing Spell Check API | Microsoft Azure," n.d.) on the corpus . A few examples for the corrections made are "uncomplete" to "incomplete", "some times" to "sometimes", "work shop" to "workshop", "people who came their without any Java knowledge" to "people who came there without any Java knowledge" and "jawa" to "java".

5. Annotation Scheme

The dataset was annotated from scratch by 4 undergraduate Computer Science students. The resulting annotations were then inspected again, and corrected (if needed). Borderline cases were resolved collaboratively by annotators.

This dataset consists of many opinionated responses. Each of those responses focuses their opinion towards a target entity or an aspect of an entity, which is called an opinion target. The phrase that carries the opinion is called an opinion expression.

We used a novel way of annotating student feedback. This was required mainly because of the nature of the data. In previous work (Luo, Liu, & Litman, 2016), data related to student feedback only had opinion targets in them whereas the positive / negative expressions were in the prompt itself. The following scheme was created to annotate the student course feedback corpus that contains responses with both opinion targets and positive/ negative expressions.

5.1 Basic Annotation Rule

We annotated important points in the response as opinion targets. Then we annotated the phrase that contains the student's opinion towards the opinion target as a positive opinion expression or a negative opinion expression, considering the polarity of the expression. For example, consider the sentence "Lectures are really good". Here, "Lectures" is the target and "really good" is the opinion expression, which is positive**.**

5.2 Annotating Pronouns

Most of the time, students refer to important entities using pronouns.

E.g.: - 1) It is good
2) It was very good to have an in-class assignment and it motivated us.

In the first example, the word "It" refers to the lecture itself. In the second example, first "It" is not an opinion target because it refers to "having an in-class assignment", which is explicitly mentioned in the sentence. Therefore, it is not needed to annotate the first "It". The second 'it" refers to "in-class assignment". Our annotation scheme is designed to annotate these pronouns. Later they will be resolved for exact entities using co-reference resolution.

5.3 Multi-word opinion targets

Consider the sentence "I think time and weight for documentation of the project is too much". Here, opinion target is "time and weight for documentation of the project", which has a negative opinion.

5.4 Single target, single expression

Consider the sentence "Lectures were really good." Here, the target is "Lectures" and the positive expression towards it is "really good".

5.5 Single target, multiple expressions

Consider the sentence "Overall lecture session was great, well organized and very helpful". Here, the target "overall lecture session" has three positive expressions towards it.

5.6 Opinion target with both positive and negative opinions

Consider the sentence "Lecture was good but a little bit fast".

Here, the student expresses his opinions about the opinion target "lecture" where the student has two opinions towards the "lecture". The opinion expression "good" expresses a positive opinion and the opinion expression "a little bit fast" expresses a negative opinion.

5.7 Single expression, multiple opinion targets

Consider the sentence "Keeping interactions with students, asking questions, giving in class activities and discussing them within the class were greatly helpful for me to develop my oop skills".

A positive opinion is expressed here for all the following aspects/ targets of the lecture: "keeping interactions", "asking questions", "giving in class activities".

5.8 Ambiguity about which opinion target to take

To resolve the matters in ambiguity while annotating the dataset, we had to come up with a hierarchy based on the entity-aspect relationship to select the best suitable target to annotate. Figure 1 explains this hierarchy.

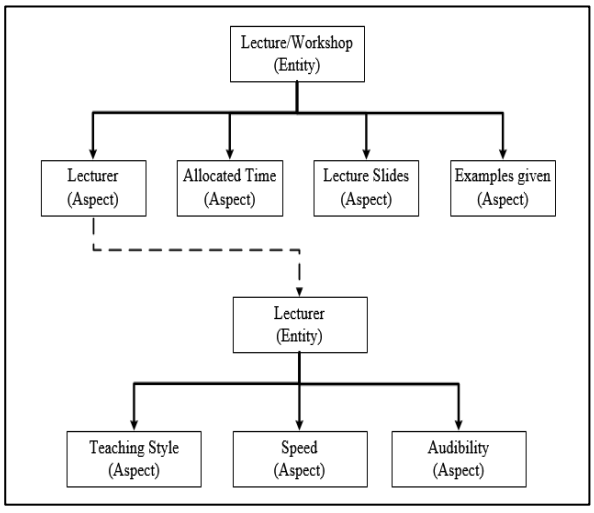

Figure 1: Example opinion target hierarchy

Aspect of one entity-aspect relationship can become the entity of another entity-aspect relationship, forming a multi-level hierarchy as shown in the Figure 1, where "Lecturer" is identified as both an aspect and an entity.

If we can identify both the entity and the aspect separately, we considered annotating both of them as targets. For example, consider the sentence "The lecture is good but the

time is not enough". Here we annotated both "lecture" (entity) and "time" (aspect) as opinion targets.

But in the cases where we cannot separate the entity and the aspect, we only annotated the higher valued target in the hierarchy, which is the target.

For example, consider the sentence "Both lecturers did a great job on delivering the subject matter." Here, out of two opinion targets that can be identified: "Both lecturers" (entity), and "delivering the subject matter" (aspect), it was difficult to decide on which target the opinion was focused on. To resolve this matter, we used the above hierarchy to prioritize the targets and annotated with the higher valued target. Since "delivering the subject matter" is done by lecturers, it becomes an aspect of "Both lecturers". Therefore "Both lecturers" is in a higher level of the hierarchy and has a higher priority. So accordingly, we annotated "Both lecturers" as the target.

We annotated the dataset manually using the above-mentioned method. This annotation scheme first identifies sentences or phrases with opinions and then marks the opinion target. Finally, the annotated phrases are converted into BIO tags. Since we annotated both the targets and the expressions, we used the following notation.

- B-T: Beginning of the Target
- I-T: Inside of the Target
- B-PO: Beginning of the Positive Opinion
- I-PO: Inside of the Positive Opinion
- B-NO: Beginning of the Negative Opinion
- I-NO: Inside of the Negative Opinion

For example, the sentence "Lectures were really good" was annotated as shown in Table 2.

Lectures	were	really	good.
B-T	O	B-PO	I-PO

Table 2: BIO tagging example

5.9 Statistics of the annotated corpus

Statistics of the student course feedback corpus annotations are given in Table 3. Out of 2,395 sentences in the corpus, 1,780 sentences contain 2,125 opinion targets.

Total no. of Opinion targets	2,125
Avg. opinion targets per lecture/ workshop	106
Max. opinion targets in a lecture/ workshop	199
Min. opinion targets in a lecture/ workshop	9

Table 3: Statistics of Annotations

Statistics of unique targets in the annotated data set are given in Table 4. Only 29 out of 106 targets were unique from an average course feedback set.

Avg. unique targets per lecture/workshop	29
Max. unique targets per lecture/workshop	54
Min. unique targets per lecture/workshop	4

Table 4: Statistics of unique targets

Distribution of the opinion targets in the student course feedback corpus annotation is depicted in Figure 2.

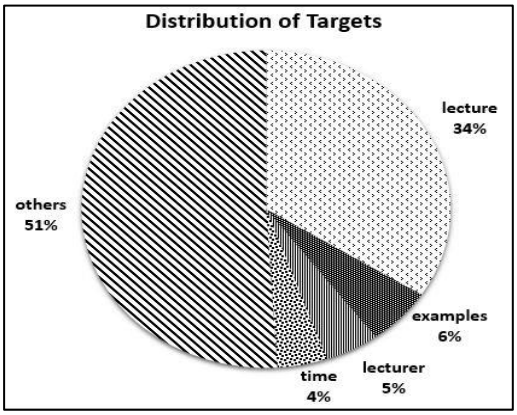

Figure 2: Distribution of opinion targets

Most frequently mentioned opinion targets by the students are "lecture", "examples", "lecturer" and "time". "Others" category which cannot be resolved to any of the hierarchies thus not shown in Figure 1, contains all medium and low frequent opinion targets, where more than 50% of the opinion targets are mentioned around 5-20 times. Least frequently mentioned opinion targets are "homework", "book", "board", "UML", "fundamentals of TDD", "log4j", "Junit", "threads" and "databases". Most of these are specific to a single workshop/lecture. These low frequent opinion targets are difficult to identify using supervised learning approaches. If we considered a fixed set of aspect categories rather than annotating each different target, less frequent targets will not be extracted. It was the reason to follow a novel annotation scheme as described in section 5.

The reliability of the annotations was verified using inter-annotator reliability from which the percent agreement was 89.2%, and the Kappa (Cohen, 1960) was 0.616. The percent agreement was calculated by token wise considering whether it is within an opinion target or an opinion expression.

The high ambiguity of the dataset resulted in this lower Kappa statistic. For example, consider the sentence "The first 7 lectures have been really good". In this example, one annotator has annotated "first 7 lectures", while the other one has annotated only "lectures" as the target. In general, most of these conflicts are due to the ambiguity of aspects and entities. Since we are annotating targets, they could be entities as well as nouns. In some cases, for annotators it is a difficult task to decide whether to annotate noun or the entity. In some cases, annotating both could be appropriate. This is because the ambiguity and noise of general feedback. Therefore, inter-annotator agreement and Kappa

value is quite low compared to reflective prompt-based annotations.

6. Conclusion

In this work, we have focused on annotating a corpus of general student feedback, which contains more noise and complex data compared to feedback collected through reflective prompts. However, unlike traditional reflective prompts, general feedback contains more useful information. Therefore, it is important to analyse general feedback. Annotating general feedback is a challenging task because of the ambiguity and noise. Here we proposed a simple annotation scheme with clarity to annotate general feedback for sentiment analysis. We have addressed the ambiguity and noise with various solutions in this dataset of 20 feedback sets with 973 responses.

The main aim of the corpus is to be used for opinion target extraction, polarity detection, and summarization of general student feedback. Since this dataset contains both positive and negative sentiments for a target, this dataset can be used to extract both positive and negative sentiments towards different aspects of a course unlike in other related corpora. We believe that this corpus can be used to improve and explore features to be used for target extraction using sequence labelling.

7. Bibliographical References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Brody, S., & Elhadad, N. (2010). An unsupervised aspect-sentiment model for online reviews. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 804–812.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Erdogan, H. (2010). Sequence labeling: Generative and discriminative approaches hidden markov models, conditional random field and structured svm. *Proceedings of the Tutorial at International Conf. Machine Learning and Applications*, 726–733.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 168–177.

Lafferty, J., McCallum, A., & Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of Intl. Conf. on Machine Learningm*, 282-289.

Luo, W., & Litman, D. (2015). Summarizing Student Responses to Reflection Prompts. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1955–1960.

Luo, W., Liu, F., & Litman, D. (2016). An Improved Phrase-based Approach to Annotating and Summarizing Student Course Responses. *Proceedings of the 26th International Conference on Computational Linguistics (COLING-16)*, 53–63.

Luo, W., Liu, F., Liu, Z., & Litman, D. (2016). Automatic Summarization of Student Course Feedback. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 80–85.

Microsoft Cognitive Services—Bing Spell Check API | Microsoft Azure. (n.d.). Retrieved September 9, 2017, from https://azure.microsoft.com/en-us/services/cognitive-services/spell-check/

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., Al-Ayyoub,M., Zhao, et al. (2016). SemEval-2016 Task 5: Aspect Based Sentiment Analysis. *Proceedings of the 10th International Workshop on Semantic Evaluation*, 19–30.

Saif, H., Fernandez, M., He, Y., & Alani, H. (2013). Evaluation datasets for Twitter sentiment analysis a survey and a new dataset, the STS-Gold. *Proceedings of the CEUR Workshop*, *1096*, 9–21.

Sang, E. F., & Veenstra, J. (1999). Representing text chunks. *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics*, 173–179.

Turian, J., Ratinov, L., Bengio, Y., & Turian, J. (2010). Word Representations: A Simple and General Method for Semi-supervised Learning. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 384–394.

Welch, C., Mihalcea, R., Street, H., & Arbor, A. (2016). Targeted Sentiment to Understand Student Comments. *Proceedings of the 26th International Conference on Computational Linguistics*, (1), 2471–2481.

Zhang, Z., & Lan, M. (2015). Ecnu: Extracting effective features from multiple sequential sentences for target-dependent sentiment analysis in reviews. *Proceedings of the 9th International Workshop on Semantic Evaluation* (SemEval 2015), 736-741.