Towards Automatic Classification of Teacher Feedback on Student Writing

Gary Cheng, Julia Chen, Dennis Foung, Vincent Lam, and Michael Tom

***Abstract*—This paper reports and discusses the results of a study aimed at automatically categorising teacher feedback on student writing. A total of 3412 teachers' written comments on 90 students' draft essays were collected from an EFL course offered by a Hong Kong university during the first semester of 2016/17. The data were primarily used to design and implement an automated tool to classify teachers' comments with respect to a taxonomy of their characteristics. The findings of this study show that the performance of the automated tool is comparable to that of human annotators, suggesting the feasibility of using the automatic approach to identify and analyse different types of teacher feedback. This study can contribute to future research into the investigation of the impact of teacher feedback on student writing in a big data world.**

***Index Terms*—Teacher feedback, draft essay, automatic classification, EFL writing.**

I. Introduction

There has been increasing recognition of the need to help students focus on the process of writing rather than merely on the final product [1], [2]. The process-oriented approach is therefore widely used in the writing class where students, especially those of English as Foreign Language (EFL), are engaged in various stages of the writing process (e.g. pre-writing, drafting and revising). Throughout this process, students will often produce multiple drafts of their essay and receive the teacher's comments on each draft. They will also need to consider all comments on their draft, diagnose its problems, make revisions and improvements as needed, and create the next version of their essay [3].

Feedback plays a critical role in supporting students' revisions of their writing [3]-[5]. Previous research showed that novice writers tend to make surface changes (e.g. change of spelling, tense or punctuation) while experienced writers typically make text-based changes (e.g. meaning, organisation or coherence) [6], [7]. There was also evidence indicating that the success of student revision is attributed to certain types of teacher feedback such as text-specific comments [3] and comments identifying problems [4].

However, findings about the effects of feedback on draft revision should be taken with caution since they were drawn from a limited number of studies and student cases [8], [9]. Moreover, the reported studies and cases were largely conducted amongst EFL students in English-speaking countries. The results may not be generalised to other populations where students are living in their hometown such as Chinese-speaking students in Hong Kong [10].

To address the limitations of previous studies, a project was initiated to develop, implement and evaluate automatic tracking of student responses to teacher feedback in draft revision. The present study, which is part of the project, aimed to develop an automated tool to classify teacher feedback on students' draft essays and provide statistics about the use of different types of teacher feedback on the drafts. The automated tool makes use of both the syntactic and semantic structures of written comments to identify their types and generate relevant statistical information.

This study can contribute to illustrating the feasibility of automatic classification of teacher feedback. It can also open up the opportunity for teachers to be promptly informed about their use of feedback types. Moreover, it can play a key part in supporting automatic analysis of the relationship between teachers' comments and students' revisions in future research.

II. Study Context

This study took place at the English Language Centre of a Hong Kong university in the first semester of the academic year 2016/17. Ninety-two undergraduate students (30 males and 62 females) taking a 13-week, credit-bearing English language enhancement course entitled 'Advanced English for University Studies' (AEUS) participated in the study. As part of the course assessment, students were required to submit two academic position argument essays on a topic of their own choice. The first was a 600-word draft, and the second was a polished, final essay on the same topic of 1200 words.

TABLE I: Details of the Participating Class Groups

Class Group ID	Programme of Study	No. of Participants	Instructor ID
A&F	Accounting & Finance	20	IA
AD	Advertising Design	11	IC
MHN	Mental Health Nursing	11	IB
N1	Nursing	16	IA
N2	Nursing	15	IB
P	Physiotherapy	19	IB

Manuscript received April 30, 2017; revised September 23, 2017. This work was supported in part by the Hong Kong SAR Government under General Research Fund (GRF no. 18608816).

Gary Cheng is with the Department of Mathematics and Information Technology, The Education University of Hong Kong, Tai Po, Hong Kong (e-mail: chengks@eduhk.hk).

Julia Chen is with the Educational Development Centre, The Hong Kong Polytechnic University, Hung Hom, Hong Kong (e-mail: julia.chen@polyu.edu.hk).

Dennis Foung, Vincent Lam, and Michael Tom are with the English Language Centre, The Hong Kong Polytechnic University, Hung Hom, Hong Kong (e-mail: dennis.foung@polyu.edu.hk, vincent.wk.lam@polyu.edu.hk, michael.tom@polyu.edu.hk).

doi: 10.18178/ijiet.2018.8.5.1060

Participants' ages ranged from 17 to 21 years (M=18.15 and SD=0.94). They came from six class groups and five academic disciplines: Advertising Design, Accounting and Finance, Mental Health Nursing, Nursing, and Physiotherapy. In addition to the student participants, three English instructors teaching the participating classes were involved in the study. Table I shows the details of the participating class groups.

III. Taxonomy of Feedback Types

Straub [11] proposed six categories of teacher feedback to characterise the ways that teachers frame their comments. The categories include praise, criticism, imperative, advice, closed question and open question. The most controlling feedback types are criticism and imperative because they request changes in a strong authoritative mode. Advice is less controlling than criticism and imperative as it usually offers suggestions using qualifiers and conditionals. Praise reflects the teacher's values, but it does not imply any changes. Closed question requests an evaluation or indirectly ask students to consider changes, while open question gives students hints to figure out problems on their own. Given its high relevance to the context of the present study, Straub's taxonomy of feedback types [11] was employed in this study. Details of the taxonomy can be found in Table II.

TABLE II: Taxonomy of Feedback Types

Code	Category	Description	Example
T1	Praise	Positive comments, non-controlling	Well written
T2	Criticism	Negative comments or evaluations, authoritative	Confusing
T3	Imperative	Comments that tell the student to do or change something, usually starting with a verb in imperative form	Be consistent
T4	Advice	Suggestive comments often in conditional mode	Maybe you could add some details here
T5	Closed question	Questions that either get a 'yes' or 'no' as answer, or else a simple one-word answer	Do you think you have given an adequate evaluation?
T6	Open question	Questions that require more than a 'yes' or 'no' answer, often starting with 'what', 'where', 'why', 'who, 'when' and 'how'	What does this mean?

IV. Automatic Classification Method

The automatic classification method comprises three main processes. First, a data set of manually annotated teacher feedback was compiled. Second, key features (i.e. syntactic, heuristic and semantic) signifying a specific feedback category are extracted. Finally, a new comment is systematically classified into a feedback category based on their similarity in features. Details of the steps are given below.

A. Compiling a Data Set

A total of 3412 teachers' written comments on 90 students' draft essays were collected. The total size of the comments in the data set is 20478 words, resulting in an average length of 6 words per comment.

Every comment in the data set was manually annotated by two researchers with reference to Straub's taxonomy of feedback types [11]. Discrepancies in results between the researchers were discussed to reach a consensus on the annotating standards. The basic unit of annotation was a single sentence, and each unit was labelled with one feedback category only. Fig. 1 shows a sample text and its associated comments marked up with feedback categories.

Student text:
The Yulin Dog Meat festival provokes arguments and flames between dog lovers and dog eaters.

Teacher feedback:
What is the Yulin Dog Meat festival? **[T6]** Since you are writing in English, you cannot assume your reader knows what this is. **[T2]** Explain clearly first. **[T3]**

Fig. 1. A sample student text with annotated teacher comments.

B. Identifying Syntactic and Heuristic Rules

Each feedback category may possess a set of distinctive syntactic rules and heuristics (see Table III). Syntactic and heuristic analyses were performed on the data set to identify rules that could be used for classification of teacher feedback. This process comprises the following steps:

1) Extract all comments labelled with the same category from the data set and put them into the same 'bag of comments'.
2) Use part-of-speech (POS) tagger provided by Natural Language Toolkit (NLTK) [12] to assign a POS tag (e.g. NN: noun, VB: verb in base form, JJ: adjective, and etc) to every word in a sentence. A sequence of POS tags can be understood as a syntactic rule.
3) Identify a set of syntactic rules that can distinguish a feedback category from another.
4) Compile a set of keywords and keyword sequences that are most likely found in a feedback category. Each element from the set corresponds to a specific heuristic rule.
5) Reduce and merge common or similar rules to maintain a minimum set of distinctive syntactic and heuristic rules for each category.

C. Extracting Semantic Features

Apart from syntactic and heuristic rules, it is necessary to extract semantic features that best discriminate between feedback categories. Specifically, the following steps were performed:

1) Tokenise each 'bag of comments' (i.e. T1 to T6) into a 'bag of words' using built-in functions given by Natural Language Toolkit (NLTK) [12].
2) Filter out non-essential words like empty strings, duplicates and stop words (e.g. a, the, and) from each 'bag of words'.
3) Transform every word into its root form. For example, 'watching' and 'watched' could be converted into the

same root form 'watch'.

4) Calculate the weight of each word in a 'bag of words' using TF-IDF normalisation method [13]. The TF-IDF method is a product of Term Frequency (TF) and Inverse Document Frequency (IDF) whereby rare words get more weight while common words can get less.
5) Construct a word-by-category matrix where each row corresponds to a word and each column refers to a feedback category. Each cell(x, y) in the matrix indicates the weight of a word x in a feedback category y. This matrix is sometimes large and sparse.
6) Apply Singular Value Decomposition (SVD) to transform the word-by-category matrix into a lower-order space based on Latent Semantic Analysis (LSA) [14]. The transformation is a form of factor analysis that aims to reduce the matrix dimension by discarding insignificant features.

TABLE III: SOME SYNTACTIC AND HEURISTIC RULES

Code	Category	Example
T1	Praise	• Positive word (e.g. You have done *good* research on the topic)
T2	Criticism	• Negative word (e.g. The topic sentences and concluding sentences are often *problematic*) • Negation + positive word (e.g. The second paragraphs did *not* flow *well*)
T3	Imperative	• Starting with a verb (e.g. *Use* another linking word) • 'must' + verb (e.g. As an academic research essay, you *must use* peer-reviewed academic journal articles)
T4	Advice	• Starting with 'You should' (e.g. *You should* have the example earlier or should not have the example) • Starting with 'You are recommended' (e.g. *You are recommended* to demonstrate a range of citation strategies)
T5	Closed question	• Starting with 'Is' and ending with '?' (e.g. *Is* this a direct quotation*?*) • Ending with 'right?' (e.g. This source seems to be irrelevant, *right?*)
T6	Open question	• Starting with 'What', 'Where', 'Why', 'Who', 'When' and 'How' and ending with '?' (e.g. *Why* would you use "in addition" at the beginning of your concluding sentence*?*)

D. Classifying a New Comment

A new comment is first processed at the sentence level by the syntactic and heuristic approach. Every sentence will be tagged by the POS tagger and will then be matched against existing syntactic and heuristic rules to identify its category. If this approach fails to return a category, the semantic approach will be applied next to the comment.

In the semantic approach, a sentence is first encoded by a vector of word weights. The vector's cosine similarity to different feedback categories will be calculated. Cosine similarity measures the cosine of the angle between two vectors. The similarity value is 1 if the two vectors are identical while it is 0 if the two vectors are orthogonal (i.e. they are completely different). A sentence will be classified into a specific category where the similarity between them is the highest among other categories. Fig. 2 illustrates the process of how to automatically classify a teacher comment into a feedback category.

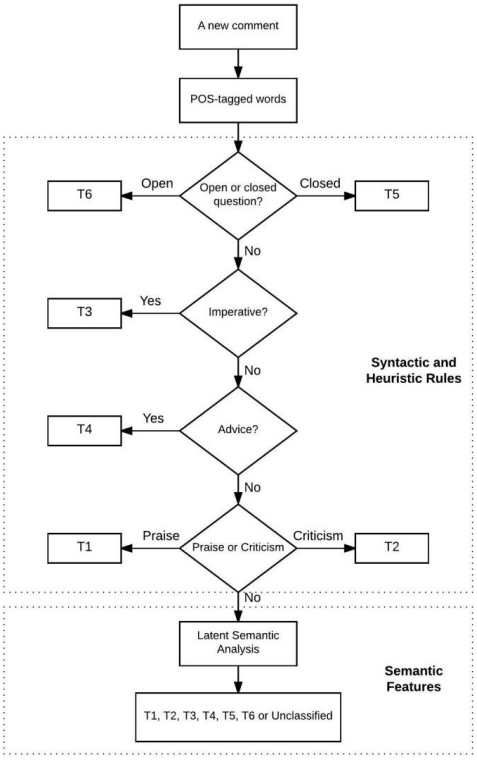

Fig. 2. The process of automatic classification.

V. RESULTS AND DISCUSSION

A. Distribution of Feedback Categories

The overall distribution of manually annotated feedback categories in the data set is shown in Table IV. From the table, it can be seen that the top three most frequently used feedback categories are T2 (criticism, 52.2%), T3 (imperative, 15.9%) and T6 (open question, 11.0%). The results suggest that teachers were inclined to use the most controlling feedback categories as a means to call for students to make revisions.

When taking a closer look at the distribution of feedback categories in each class group, instructors appeared to have different use patterns of feedback categories as evident in Table V. Even though they all overwhelmingly used the most controlling feedback categories (T2 or T3), instructor IC used more imperatives (T3) than criticisms (T2) but other instructors did the opposite. Moreover, there was an apparent difference in the usage percentage of praise (T1) between instructor IB and other instructors.

TABLE IV: DISTRIBUTION OF FEEDBACK CATEGORIES IN THE DATA SET

Feedback Category	Count	Percentage (%)
T1	197	5.8
T2	1782	52.2
T3	543	15.9
T4	263	7.7
T5	250	7.3
T6	377	11.0

Not only could the use pattern of feedback categories be influenced by the quality of student essay, but it could also likely be determined by the teacher's preference and habit. An automatic analysis of the distribution of feedback

categories would be helpful for teachers to reflect on their own choice of feedback types. Based on the classification results and statistical data, they could use their professional judgement to decide whether or not to adjust their feedback practice.

TABLE V: PERCENTAGE OF FEEDBACK CATEGORIES IN DIFFERENT CLASS GROUPS

Class Group ID	Instructor ID	Percentage (%)					
		T1	T2	T3	T4	T5	T6
A&F	IA	0.9	59.5	14.9	6.9	7.0	10.8
N1		0.7	56.4	16.9	10.5	6.1	9.3
P	IB	19.9	51.6	7.4	6.9	6.9	7.4
MHN		13.9	54.6	7.7	6.2	7.0	10.6
N2		16.0	50.8	6.9	10.8	6.6	8.8
AD	IC	0.7	27.5	40.4	3.7	11.3	16.4

B. Accuracy of the Automatic Classification Method

The data set were randomly split into two equal-sized groups: Group 1 and Group 2. Group 1 was first used as training data for feature extraction and Group 2 as testing data for performance evaluation. This process was performed one more time, with Group 2 for training and Group 1 for testing. Accuracy, which is defined as the proportion of machine classifications that agree with manual classifications [15], was calculated on testing data to evaluate the effectiveness of the automatic classification method.

Table VI presents the confusion matrix generated from the two-fold evaluation on the data set. It is a table containing information about manual and machine classifications. The elements of the main diagonal represent the number of comments for which the category identified by machine is the same as the actual category identified by human, while the remaining elements are those that are mis-classified by the machine. Apart from the six feedback categories (i.e. T1 to T6), one extra category namely 'U' was created to represent the unclassified result generated by machine. As Table VI shows, the accuracy of the automatic classification method ranges from 96.6% to 100%, indicating that the overall performance of the proposed method in classifying teacher feedback is very good.

TABLE VI: CONFUSION MATRIX AND ACCURACY

		Machine Classification							Accuracy
		U	T1	T2	T3	T4	T5	T6	
Human Classification	T1	0	*197*	0	0	0	0	0	100%
	T2	7	46	*1722*	4	3	0	0	96.6%
	T3	0	8	2	*533*	0	0	0	98.2%
	T4	0	5	2	1	*255*	0	0	97.0%
	T5	0	0	0	0	0	*244*	6	97.6%
	T6	0	0	0	0	0	9	*368*	97.6%

VI. CONCLUSION AND IMPLICATIONS

In this study, an automated tool was designed and implemented to identify types of teacher feedback on students' draft essays. Based on Straub's taxonomy of feedback types, the automated tool uses both syntactic and semantic approaches to classify a teacher's comment into one of the six categories (i.e. praise, criticism, imperative, advice, closed question, and open question). The findings of the study show that the automated tool performed very well in terms of the accuracy of classifying teacher feedback. They also indicate that instructors tended to overwhelmingly use the most controlling feedback categories (i.e. criticism and imperative) but the use pattern of feedback categories varied across instructors.

The implications of the study are that by providing instant analysis of their use pattern of feedback types, teachers would better reflect on their feedback practice and then make appropriate changes. Perhaps more importantly, the proposed method can become a key part of future research aiming at automatically analysing the relationship between teacher feedback and student revision. This would in turn contribute to a fuller understanding of the impact of teacher feedback on student writing.

REFERENCES

[1] L. Flower and J. R. Hayes, "A cognitive process theory of writing," *College Composition and Communication*, vol. 32, no. 4, pp. 365-387, 1981.

[2] K. Hyland, *Second Language Writing*, New York: Cambridge University Press, 2003.

[3] D. R. Ferris, "The influence of teacher commentary on student revision," *TESOL Quarterly*, vol. 31, no. 2, pp. 315-339, 1997.

[4] S. M. Conrad and L. M. Goldstein, "ESL student revision after teacher-written comments: Text, contexts, and individuals," *Journal of Second Language Writing, vol. 8, no. 2, pp.* 147-179, 1999.

[5] G, Cheng, "The impact of online automated feedback on students' reflective journal writing in an EFL course," *The Internet and Higher Education*, vol. 34, pp. 18-27, 2017.

[6] L. Faigley and S. Witte, "Analyzing revision," *College Composition and Communication*, vol. 32, no. 4, pp. 400-414, 1981.

[7] J. Fitzgerald, "Research on revision in writing," *Review of Educational Research*, vol. 57, no. 4, pp. 481-506, 1987

[8] D. R. Ferris, *Response to Student Writing: Implications for Second Language Students*, Mahwah, NJ: Lawrence Erlbaum, 2003.

[9] C.-Y. Chiu and S. J. Savignon, "Writing to mean: Computer-mediated feedback in online tutoring of multidraft compositions," *CALICO Journal*, vol. 24, no. 1, pp. 97-114, 2006.

[10] J. Chen and L. Hamp-Lyons, "Effective feedback on student writing," *Quality in Teaching and Learning in Higher Education: A Collection of Referred Papers from the first Conference*, Hong Kong: Hong Kong Polytechnic University, 1999, pp. 113-120.

[11] R. Straub, "Students' reactions to teacher comments: An exploratory study," *Research in the Teaching of English*, vol. 31, no. 1, pp. 91-119, 1997.

[12] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*, O'Reilly Media, 2009.

[13] H. C. Wu, R. W. P. Luk, K. F. Wong, and K. L. Kwok, "Interpreting TF-IDF term weights as making relevance decisions," *ACM Transactions on Information Systems*, vol. 26, no. 3, pp. 1-37, 2008.

[14] T. K. Landauer, P. W. Foltz, and D. Laham, "Introduction to latent semantic analysis," *Discourse Processes*, vol. 25, pp. 259-284, 1998.

[15] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed. Burlington, MA: Morgan Kaufmann, 2011.

Gary Cheng is an assistant professor of the Department of Mathematics and Information Technology at the Education University of Hong Kong. His research interests include e-learning, learning management systems, electronic portfolio, automated systems for teaching and learning, and learning analytics.

Julia Chen is director of the Educational Development Centre at the Hong Kong Polytechnic University. Previously she was an associate director of the English Language Centre, coordinating 4YC language & communication requirement subjects, materials development and QA; and was the chair of the Faculty of Humanities' Learning and Teaching Committee.

Dennis Foung is a language instructor at the English Language Centre of the Hong Kong Polytechnic University. His professional interests include learning analytics, computer assisted language learning, writing in the disciplines or writing across the curriculum, and classroom discourse.

Vincent Lam is a language instructor at the English Language Centre of the Hong Kong Polytechnic University. His professional interests include curriculum design and leadership building in education, teacher preparedness for post-modern education, formulaic sequences on second language fluency, and teaching and learning English through drama.

Michael Tom is a language instructor at the English Language Centre of the Hong Kong Polytechnic University. His professional interests include task-based language teaching, English for academic purposes, learner engagement and motivation, corrective feedback, English language teaching material development, educational technology, and e-learning.