Proceedings of the Fifth Conference on
Human Computation and Crowdsourcing
(HCOMP 2017)

Supporting ESL Writing by Prompting Crowdsourced Structural Feedback

Yi-Ching Huang,[1] **Jiunn-Chia Huang,**[2] **Hao-Chuan Wang,**[3] **Jane Yung-jen Hsu**[2,4]
[1]Graduate Institute of Networking and Multimedia, National Taiwan University
[2]Department of Computer Science and Information Engineering, National Taiwan University
[3]Department of Computer Science, National Tsing Hua University
[4]Intel-NTU Connected Context Computing Center
d00944010@csie.ntu.edu.tw, jiunnhuang@ntu.edu.tw, haochuan@cs.nthu.edu.tw, yjhsu@csie.ntu.edu.tw

Abstract

Writing is challenging, especially for non-native speakers. To support English as a Second Language (ESL) writing, we propose StructFeed, which allows native speakers to annotate topic sentence and relevant keywords in texts and generate writing hints based on the principle of paragraph unity. First, we compared our crowd-based method with three naïve machine learning (ML) methods and got the best performance on the identification of topic sentence and irrelevant sentence in the article. Next, we evaluated the StructFeed system with two feedback-generation mechanisms including feedback generated by one expert and by one crowd worker. The results showed that people who received feedback by StructFeed got the highest improvement after revision.

Introduction

Writing is a difficult task, especially for non-native speakers. To construct a well-structured essay, ESL writers usually take much efforts and time to write and rewrite iteratively. During the iterative process, many written compositions which have many writing issues like lack of clarity or focus, or incomplete topic development are generated by ESL learners. They need a pair of outside eyes to identify their weak spots and to suggest ways of fixing them. However, collecting high-quality feedback is challenging due to the limited pool of experts available. To support on-demand help, we require an approach that supports ESL writers to identify writing issues and suggest ways for improving their writing.

Previous studies have explored automated writing evaluation systems (AWE) for supporting second language writing. Almost all studies utilize supervised learning on large training datasets to predict the holistic score and to generate diagnostic feedback of an essay (Dikli 2006). However, those automated methods only support limited topics for writing practices due to the cost of training data collection, annotation, and processing. To enable diverse writing support, we leverage the power of native speaker to make small contributions for identifying basic writing elements like topic sentences and supporting data in the texts.

According to our observations, many ESL students with more than 10-year English learning experience still struggled with identifying topic sentence for a basic five-paragraph essay and usually failed to develop a unified essay in our pilot study. That is why the ESL writing pedagogy always starts with teaching topic sentence writing, and move on the development of paragraph and essay later (Oshima and Hogue 2013). Therefore, we propose StructFeed, a crowd-powered system that generates structural feedback for helping ESL writers recognize high-level writing issues and produce a unified article. A crowdsourcing workflow is used for allowing native-speakers to identify topic sentence and relevant keywords in an article. Next, the system will predict the location of topic sentences and irrelevant sentence by aggregating crowd annotations and then generate writing suggestions. The goal is to guide people to revise the paragraph to achieve the paragraph unity based on writing criteria.

We compare our crowd-based method with three naïve machine learning (ML) methods. The results suggest that the crowd-based method outperforms all ML methods. In addition, the new rule derived from crowd annotations outperformed all initial methods. Furthermore, we evaluate StructFeed with 18 ESL writers recruited through online postings in the community sites. A between-subject experiment was conducted to investigate how and whether people can improve their writing after receiving feedback generated by one crowd worker, one expert, or StuctFeed. The results showed that people who received writing suggestions from StructFeed achieved the best performance than other people who received writing suggestions from one crowd worker or one expert.

Related Work

In this section, we review previous work on automated writing evaluation, writing support by crowdsourcing, and external feedback.

Automated Writing Evaluation

State-of-the-art automated writing evaluation systems (AWE) for supporting second language writing utilize supervised learning on large training datasets to predict the holistic score and to generate diagnostic feedback of an essay (Dikli 2006). As an example, ETS Criterion (Burstein, Chodorow, and Leacock 2004) uses a discourse structure

Copyright © 2017, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

trained on 1,462 labeled essays, and builds a specific scoring model for each topic trained on a sample of 200250 labeled essays. Consequently, it only supports limited topics for writing practices due to the cost of training data collection, annotation, and processing. The proposed crowd-based writing framework is scalable, which supports any topic by collecting annotations on key elements of an essay (e.g., topic sentences, keywords) and providing structural feedback in alignment with the principles of writing.

Crowdsourcing Writing Task

Many researchers have explored how crowdsourcing can be applied to several types of writing tasks. Soylent (Bernstein et al. 2010) used online paid workers to perform three complex writing tasks: shortening, proofreading, and editing paragraph. MicroWriter (Teevan, Iqbal, and von Veh 2016) focus on how to use micro-tasks to support collaborative writing by providing a narrative structure as a template. Ensemble (Kim, Cheng, and Bernstein 2014) and Storia (Kim and Monroy-Hernandez 2016) allow crowd workers to generate ideas and contribute content to construct a short story. Those projects have demonstrated the viability of crowdsourcing to allow crowd workers help a user to accomplish several writing tasks. However, for long-term success, the writing skill of a user has not been developed during those process. In this paper, we focus on how to use crowd workers to facilitate writing behavior instead of accomplishing the writing task.

External Feedback

Feedback plays an essential role in a creative task like writing or design. Obtaining feedbacks from others help people recognize the gap between their intentions and others' interpretations, understand key principles, and facilitate continuous improvement. Recently, many researchers have explored the feasibility of eliciting feedbacks from peers in an online educational environment. Tinapple et al. developed a system named CritViz to enables online peer critiques of designs (Tinapple, Olson, and Sadauskas 2013). Kulkarni *et al*. created PeerStudio, an assessment platform that enables rapid feedback on in-progress work by leveraging students' peers in the massive online classes like Coursera and OpenEdX (Kulkarni, Bernstein, and Klemmer 2015). It used well-defined rubrics and scaffolding comments to help peer reviewers accomplish the review writing task.

Prior studies have shown that external feedbacks obtained from crowds help users discover problems and improve their work (Dow et al. 2012; Xu, Huang, and Bailey 2014; Luther et al. 2015). Dow *et al*. used assessment rubrics as feedbacks to help users produce better quality work. Xu *et al*. created Voyant to collect perception-oriented feedbacks including Elements, First Notice, Impression, Guidelines, and Goal from online crowds. Luther *et al*. adopted scaffolding theory to design a structure interface to guides worker to provide design suggestions. The interface embedded seven key design principles produced by two experienced authors and had successfully elicited feedback from non-experts. Inspired by Luther *et al*'s work, we apply the idea of scaffolding theory and design our crowdsourcing workflow by a standard writing theory. The task design in Topic stage and Relevance stage follow the concept of English as Second Language writing curriculum.

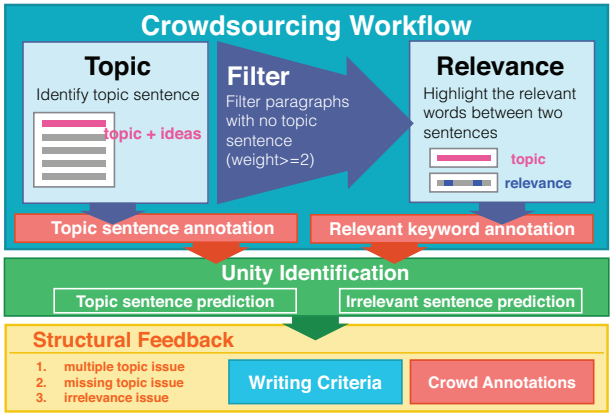

Figure 1: The overview of StructFeed. The system generated writing suggestions based on aggregated crowd annotations and writing critera.

StructFeed

StructFeed is a crowd-based system that allows a user to request, receive, and review writing feedback for recognizing and fixing structural issues of writing. Instead of providing feedback on local issues like grammatical or spelling errors, StructFeed attempts to address global issues like irrelevant ideas or missing main topic.

In this section, we introduce the design of StructFeed, and the overview of the system is depicted in Figure 1. First, we describe the essential principle of writing – paragraph unity. Next, we introduce our crowdsourcing workflow that allows crowd workers who are native speakers to examine the paragraph unity through two types of micro-tasks. Finally, we present the structural feedback with a visualization interface.

Paragraph Unity and Topic Sentence

A good essay should have a clear structure in which all elements are well organized and linked. An essay consists of introduction, body, and conclusion and each part is composed of paragraphs. A paragraph is the basic component of writing, and it is a group of related sentences that are organized to develop a single idea. It contains a topic sentence, several supporting sentences, and a concluding sentence. The topic sentence is the most important one because it indicates the main idea of a paragraph. The supporting sentences are used to provide evidence to support the main idea. The concluding sentence is used to summarize the main idea presented in the topic sentence and emphasize the impression on the readers.

A good paragraph should follow an important principle called unity. Unity is used to evaluate the quality of oneness in a paragraph or an essay. It can be achieved by the following two steps.

- All sub-points are related to one main idea.

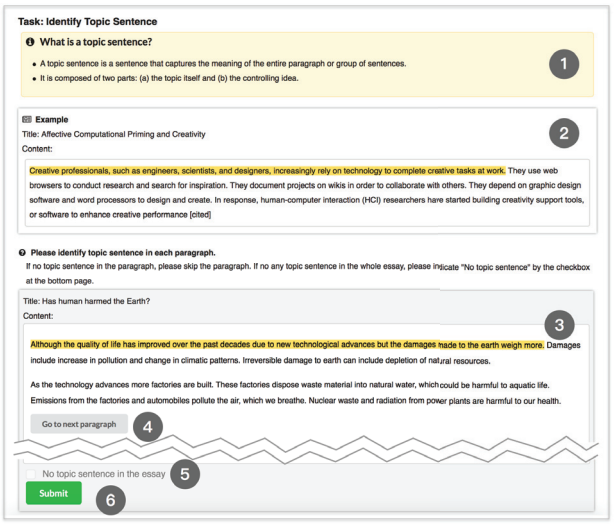

Figure 2: The crowdsourcing interface contains 1) definition of topic sentence, 2) a worked-out example, 3) working area, 4) next button, 5) check-empty button, and 6) submit button.

- No irrelevant sentence exists in the paragraph.

Crowdsourcing Workflow

The designed workflow breaks down the process of unity identification into two stages: Topic and Relevance stage.

The system dispatches micro-tasks to online crowdsourcing marketplace in both Topic and Relevance stages. There is a filter between the two stages. It aggregates results from Topic stage and passes qualified results to Relevance stage.

Topic Stage The goal of Topic stage is to examine whether all paragraphs have a topic sentence. In this stage, the system creates a task with five assignments and distribute it to distinct crowd workers. The task asks workers to mark every topic sentence in an essay. Our tool lets workers make sentence-level annotation by clicking on any part of a candidate sentence. The selected sentence will be highlighted with yellow background. The annotation can be cancelled by re-clicking.

The crowdsourcing interface in Figure 2 is designed to guide workers to accomplish the task with good quality. The interface contains a brief description of topic sentence (1), a worked-out example (2) for teaching workers how to identify topic sentence, and a working area (3). In the working area, a crowd worker can annotate a sentence with a simple click. A next button (4) is used to make workers focus one paragraph at a time; when it is clicked, the next paragraph will appear in the working area. When all the paragraphs appears, the check-empty (5) and submit button (6) will show up. In the end, a worker can submit the answer and finsih the task.

Relevance Stage The goal of Relevance stage is to determine whether every other sentence is related to the topic sentence in a paragraph.

In this stage, the system creates a task with three assignments and dispatch it to different workers. The task contains one paragraph with topic sentence labeled. The given topic sentence is determined by majority voting from the previous stage and is highlighted in yellow color. The task asks workers to locate the word which is related to the given topic sentence. Similar to the design of the previous stage, workers can make word-level annotation by clicking on any part of a candidate word. The selected sentence will be highlighted with a green background. The annotation can be canceled by re-clicking.

Filter Filter is a bridge component which aggregates all annotations generated from the Topic stage and determines which one is a topic sentence by at least two annotations labeled from different workers. Next, the Filter would choose paragraphs existing a topic sentence to pass them to the Relevance stage.

Structural Feedback and Interface

Structural feedback is designed for helping writers identify their writing issues and facilitate rewriting behaviors by prompting writing hints. The feedback consists of two elements: issue summary (1) and writing hints (2). The issue summary indicates the type of writing issue including *multiple topics issue*, *irrelevance issue*, and *missing topic issue*, and a suggested editing action (see Figure 3); the writing hints show the detailed of writing issues by a number of low-level annotations. The annotations include topic sentences, irrelevant sentences, and relevant keywords. The design of writing feedback follows Sadler's requirements for high-quality feedback (Sadler 1989).

We not only show the location but also the weight for the annotations of topic sentences and relevant keywords. The weight of annotation presents the number of agreements made from different people. The blue highlighted sentence is topic sentence. The brightness of background color indicates the number of agreement from different people. When more people annotate the same sentence as topic sentence, the background color of this sentence is much deeper than the other. In addition, the annotation of a relevant keyword is indicated by green highlighting. The brightness of background color is also determined by the number of annotations generated by workers. The red dotted underline indicates the location of an irrelevant sentence.

The two sliders (3) at the top left corner of the page are used to filter two types of annotation by different weight. By moving the slider back and forth, the writer can see the annotations with different weight appears in sequential order (see Figure 3).

Implementation

StructFeed is a Web application built in Python, Javascript, and Postgres, which has been deployed on Heroku. The two types of micro-tasks in the workflow generated as two external HITs are submitted to Amazon.com's Mechanical Turk, a popular online crowdsourcing platform. Workers who have at least 80% task approval rate are considered to perform our

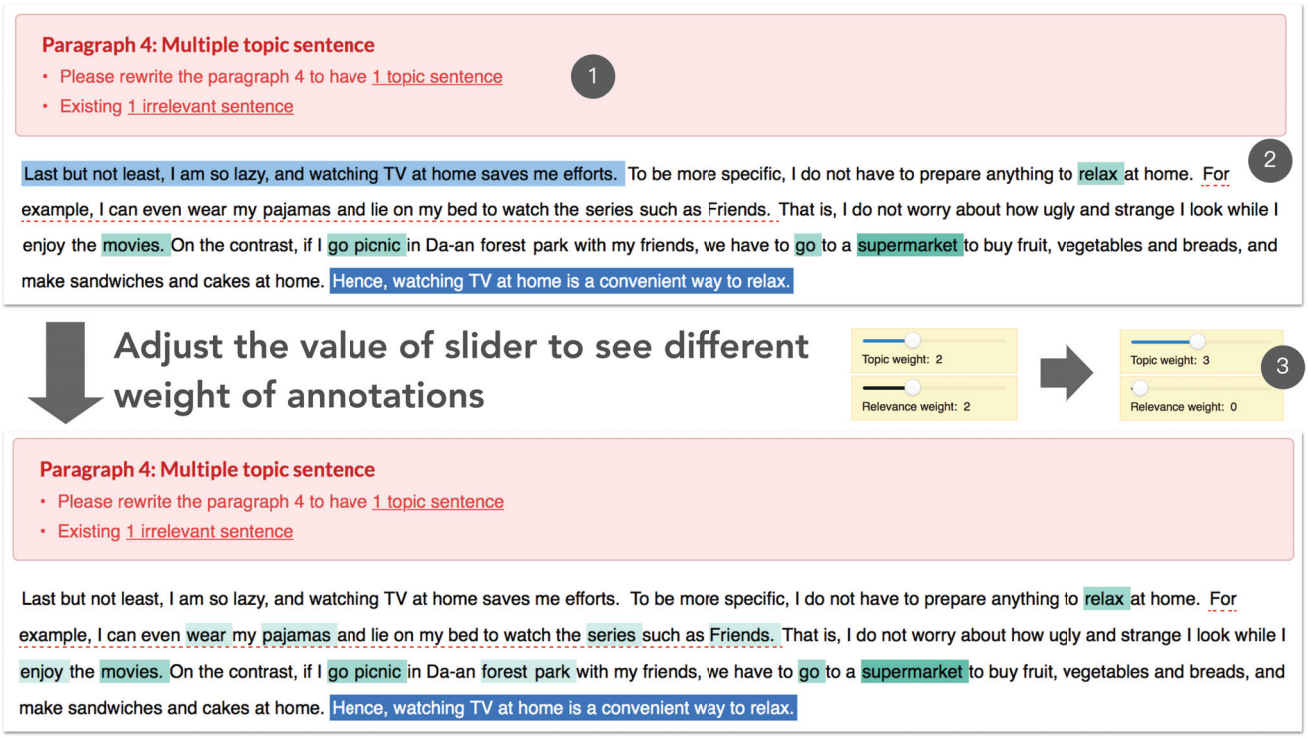

Figure 3: The feedback interface contains 1) issue summary, 2) writing hints, and 3) topic and relevance sliders. The top image shows feedback when topic weight is 2 and relevance weight is 2; the bottom image shows feedback when topic weight is 3 and relevant is 0.

tasks. Each task costs $0.05 and one worker can perform 2.5-3 tasks in a minute. The worker can get at least $7.5-$9 per hour (higher than $7.25).

Unity Identification

To support writing on any topic, StructFeed needs to work in the absence of enough data, i.e. the "cold-start problem" in the context of computer-assisted writing. Therefore, we compare our crowdsourcing approach of unity identification with three naïve machine learning (ML) methods. These ML methods are commonly used for solving problems without large amounts of labeled training data.

In this section, we evaluate all methods by calculating average precision, recall, and F1-score (combined precision and recall) of identifying topic sentence, and irrelevant sentence.

Crowdsourced-Based Methods

Topic/irrelevant sentence prediction A topic sentence is directly determined by at least two distinct crowd workers. Relevant sentence is determined by at least three distinct crowd workers who clicked the same sentence including relevant keywords. The threshold of a topic or relevant sentence is determined by empirical data discussed in the following paragraph. An irrelevant sentence is a sentence which is neither a topic sentence nor a relevant sentence.

Agreement	Precision	Recall	F1 score
1	0.459259	0.826667	0.590476
2	0.606742	0.720000	0.658537
3	0.655738	0.533333	0.588235
4	0.774194	0.320000	0.452830
5	0.733333	0.146667	0.244444

Figure 4: Results of topic sentence prediction by aggregating topic annotations from crowd workers with different threshold.

Crowd Agreement and Performance To obtain the aggregated answers, we set 2 agreements as a threshold for identifying topic sentence and 3 agreements as a threshold for identifying (ir)relevant sentence, respectively. The detail performance of topic sentence, relevant sentence, and irrelevant sentence are described in Figure 4, Figure 5, and Figure 6, respectively.

ML-Based Methods

To solve the problem lacking labeled data, we propose three naïve methods to predict topic sentence, relevant and irrelevance sentence.

Agreement	Precision	Recall	F1 score
1	0.711628	0.884393	0.788660
2	0.718085	0.780347	0.747922
3	0.722581	0.647399	0.682927
4	0.698925	0.375723	0.488722
5	0.772727	0.196532	0.313364

Figure 5: Results of relevant sentence prediction by aggregating relevant keyword annotations with different threshold.

T agree	R agree	precision	recall	f1 score
2	1	0.066667	0.031250	0.042553
2	2	0.150000	0.139535	0.144578
2	3	0.205882	0.325581	0.252252
2	4	0.158333	0.441860	0.233129
2	5	0.183007	0.651163	0.285714

Figure 6: Results of identifying irrelevant sentence by aggregating relevant keyword annotations with different threshold.

Word Similarity In our work, we choose two well-known methods for obtaining word similarity, Word2vec[1] and Wordnet[2]. Word2vec is a group of related models that are used to produce word embeddings. It is trained by skip-grams and CBOW (continuous bag-of-words) and turns words into a vector; its pairwise similarity is got by cosine distance. Wordnet is a corpus built by experts; its pairwise similarity is got by path similarity. We use these two methods to define the distance of two words and the synonym between two words.

Topic Sentence Prediction This section we purposed three kinds of method to predict topic sentence: rule-based method, TF-IDF (term frequency-inverse document frequency) and average sentence similarity. The notation s in below means a sentence, which is also a set of word w, and the set of sentences in a paragraph is noted as P. A represents the set of P, which means the whole article. The $length(s)$ represents the total word count in sentence s. The total relation between these sets is $w \in s \in P \in A$.

- Rule-based method: A paragraph usually begins with a topic sentence (Kaplan 1966). We adopted the first sentence rule in the rule-based method.
- TF-IDF: A topic sentence is a sentence that identifies the main idea in a paragraph. Each paragraph should have a different main idea in the standard essay writing. In other words, a topic sentence can be regarded as a sentence that contains the most number of keywords in the paragraph. Therefore, we adopted the concept of TF-IDF to extract keywords in the paragraph. Term frequency is calculated by sentence and inverse document frequency is calculated by paragraph. For term calculation, we aggregated the results of words with a high similarity. This method is used because with too few articles, and the same term may appear too few to find it second times. Sentence with highest average TF-IDF would be chosen to be the topic sentence in a paragraph.

$$\max_{s \in P} \sum_{w \in s} tfidf_{w,P} \Big/ length(s)$$

$$tfidf_{w,P} = \frac{n_{w,s}}{\sum_{s \in P} \sum_{w' \in s} n_{w',s}} \times log \frac{|P|}{|\{P \in A : w \in P\}|}$$

- Average sentence similarity (ASS): If there exists a topic sentence, it should have smallest average distance to other sentences in the paragraph. This method based on the corpus and calculate a pairwise distance of words between sentences in a paragraph, and then average the sum by the word aggregation. Sentence with smallest average distance sum would be chosen to be the topic sentence of a paragraph.

$$\min_{s \in P} \sum_{w_i \in s} \sum_{w_j \in P \setminus s} distance(w_i, w_j) \Big/ length(s)$$

Irrelevant Sentence Prediction Based on the outcome of topic sentence prediction, we predicted irrelevant sentences by calculating the similarity of a sentence with a given topic sentence based on two kinds of corpus mentioned above. For Word2vec, we used "cosine similarity"; for Wordnet, we used "path similarity." Then, we used leave-one-out validation to train PLA (perceptron learning algorithm) for (ir)relevant sentence identification. The ground truths are annotated by two experts (as described in the following paragraphs).

Evaluation

To evaluate these methods, we recruited 15 participants who are all non-native speakers to write an essay in 30 minutes and crowdsourced those essays to obtain annotations from crowd workers. We compared precision, recall, and F1-score for evaluating the performance of prediction of topic sentences and irrelevant sentence.

The precision is the number of correct annotations divided by the number of all collected annotations. The recall is the percent of all correct annotations that are collected from an essay. The F-score that combines precision and recall is also used to evaluate the effectiveness of retrieved sentence.

Ground-Truth Data We recruited two experts with 5+ ESL teaching and training experience to construct the gold standard annotations for 15 essays. Both experts have Ph.D. degree, and one's major is Applied Linguistics, and the other's is English Education. They were asked to annotate topic sentence, relevant keywords, and irrelevant sentence independently. The topic annotation (Kappa $k = 0.98$, $p < .0001$) and relevant keyword annotations (Kappa $k = 0.92$,

[1]http://deeplearning4j.org/word2vec
[2]https://wordnet.princeton.edu

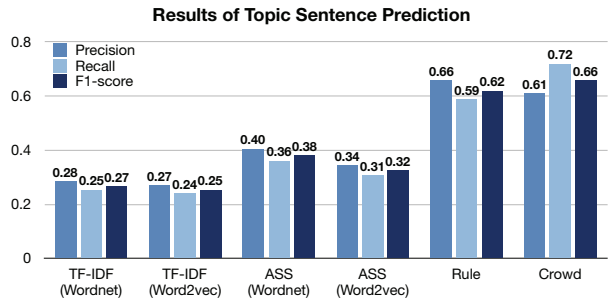

Figure 7: Results of topic sentence identification from ML-methods and our crowd-based method

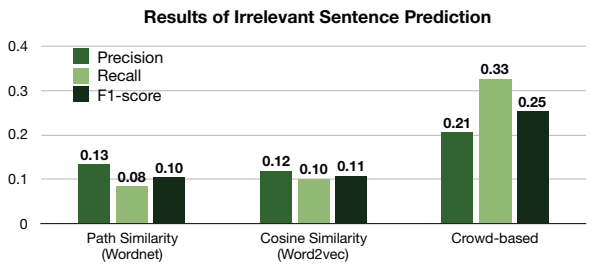

Figure 8: Results of irrelevant sentence identification from ML-methods and our crowd-based method

$p < .0001$) created by two experts had high consistency. The gold standard annotations are the union of two experts' results.

According to our observations, the two experts followed consistent principles to annotate relevant keywords.

- identify supporting data with high relevance to the topic sentence
- identify the synonym appearing in the sentence with an argument but ignore simply repeating keywords
- identify chunks with a specific relation like cause-and-effect, etc.

Crowd Annotations We crowdsourced 15 essays using our workflow in Amazon Mechanical Turk. 106 distinct workers were recruited for identifying topic sentence and relevant keywords. 55 workers completed 75 topic tasks (15 HITs with 5 assignments) and 51 workers completed 445 relevance tasks (89 HITs with 5 assignments). In total, there were 336 topic annotations and 1923 relevance annotations created in the workflow. The total cost is $26.

Topic Sentence Prediction Figure 7 shows the results of topic sentence prediction for three ML-based methods and the crowd-based method. The best performance is the crowd-based method (agreement=2). Its precision, recall and F1-score 0.61, 0.72, and 0.66, respectively. The worst result comes from TF-IDF with Wordnet, its precision, recall, and F1-score is 0.28, 0.25, and 0.27, respectively. The rule-based method (all-first) is slightly worse than the crowd-based method. In detail, it has higher precision but lower recall than the crowd-based method.

Irrelevant Sentence Prediction Figure 8 shows the results of irrelevant sentence prediction. The crowd-based method (agreement=3) outperforms other methods on irrelevance sentence prediction. Its precision, recall, and F1-score is 0.21, 0.33, and 0.25, respectively. While the result of the worst one, the similarity with Wordnet, is 0.13, 0.08, and 0.10, respectively. The recall of Crowd is three times higher than similarity-based methods.

Field Deployment Study

To evaluate our system, we conducted a field deployment study for understanding how and whether StructFeed benefits ESL learners and helps them improve the quality of writing. In this study, 18 participants aged 19-35 years (56% male) were recruited from online postings on social media sites such as Facebook and Bulletin Board System. Each participant was self-motivated and had a common goal of practicing his/her writing skill. No compensation was provided to participants.

Study Design

We used a between-subjects study design and separated 18 participants into three groups. Three feedback-generation mechanisms were compared: expert feedback, free-form crowd feedback, and StructFeed. The first mechanism is traditional writing feedback generated by one expert. Experts were recruited from Wordvice[3], a professional online editing, and proofreading service. We used a particular service named TOEFL writing editing. The expert will edit and provide diverse feedback on the structure, content, grammar, and word choices of the article based on the grading rubric of TOEFL iBT. The second mechanism is free-form writing feedback generated by a single crowd worker recruited from Amazon Mechanical Turk. The crowd worker is asked to provide general writing suggestions about how to improve the unity and coherence of writing. The third mechanism is structural feedback with topic annotations, relevant keywords, and hints for writing generated by StructFeed.

Tasks and Procedure

The study consist of two tasks: writing task and rewriting task. Each task was performed on a different day. First, each participant was asked to perform a writing task in 30 minutes with one of three selected topic. The three selected topics were obtained from the list of past TOEFL independent writing questions, published by Educational Testing Service (ETS). Each group of participants generated an equal number of writing with the three different topics. Next, each participant was asked to revise their writing based on feedback obtained from one of the three feedback-generation mechanisms. After each task, participants were allowed to check grammatical errors of their writing with Grammarly[4], a popular online grammar checker.

[3]http://wordvice.com

[4]https://www.grammarly.com

	Expert	Crowd	StructFeed
Time	24~48 hrs	10~30 mins	1~5 hrs
Quantity	55.44	8.11	-
Cost (USD)	$16	$2	$1.5~$1.7

Figure 9: Summary of three feedback-generation mechanisms.

	Mean of Diff-Rating	# of Equal Diff-rating	# of Decreased Diff-rating
Expert Feedback	0.29 (SD=0.43)	1	1
Crowd Feedback	0.38 (SD=0.44)	1	1
StructFeed	0.54 (SD=0.25)	0	0

Figure 10: Results of revision quality by three feedback-generation mechanisms.

Measure

To measure the quality of the original and revised articles, we recruited two experts to rate all essays based on the writing scoring rubrics of TOEFL iBT. The rating scale is from 0 to 5 with a 0.5 interval. The higher the rating, the better is the quality of writing. Two raters independently rated all essays blind to condition. Ratings from the two raters were highly correlated, showing sufficient inter-rater reliability (Pearson $r = 0.93$, $p < .0001$). In the end, we averaged the ratings of the two raters to get a quality measure for each writing. We also calculated the difference between the average rating of the original article and the revised article to get a measure of rewriting performance for each participant.

Results

In this study, we first summarized the performance of each feedback-generation mechanisms by time, quantity, and costs. Next, we compared the mean of the difference of rating between original article and revised one, the number of equal rating, and the number of decreased rating.

Time, Quantity, Costs The results show in Figure 9. We usually spent $16 to get expert feedback back in 24-48 hours, and got 55.44 comments (including grammar fix and suggestions about organization) for each article; we spent $2 to get crowd feedback back in 10-30 minutes, and got 8.11 comments; we spent $1-1.7 to get feedback by StructFeed in 1-5 hours.

Difference of Rating We compared the mean of the difference of rating between original articles and revised articles for each group. We also investigate whether all participants improve the quality of writing after receiving feedback. The results show in Figure 10. People who received StructFeed got the best performance than other mechanisms. Every participant increased the quality of writing after receiving feedback generated by StructFeed. Surprisingly, participants who received expert feedback got the worse performance. In addition, each free-form group (crowd and expert feedback) had two participants who had one equal rating and one decreased rating after revision. We will discuss this interesting phenomenon in the discussion session.

Discussion

Crowd helps develop better rules for machine

The results showed that crowd-based method outperformed all ML-based methods regarding identifying topic sentence and irrelevant sentence. According to our observation of aggregated crowd annotations, we found that the topic sentence of the introduction paragraph is usually the last sentence instead of the first one. Therefore, we modified the rule-based method to satisfy the new rule which the topic sentence is the last sentence in the first paragraph and the first sentence in the other paragraphs. The new rule derived from crowd annotations outperformed all initial methods. The precision, recall, and F1-score is 0.81, 0.72, and 0.76, respectively.

The obtained results also corresponded to the findings drawn from the interviews. Many participants reported that they would follow "four-paragraph essay template" to write a TOEFL independent essay in 30 minutes; besides, few participants who lacking knowledge about essay writing still kept putting their topic sentence at the beginning of all paragraphs. Therefore, the results suggest that crowd-based method is the better choice to analyze ESL writing which may be poor-structured and high diversity. Furthermore, the ultimate goal of proposed framework is to enable sustained crowd-machine collaboration. The writing support system and the ESL writer can gradually improve their skills together.

StructFeed not only identifies writing issues but promotes reflection

Unlike the automated methods, StructFeed not only identifies and locates topic sentence and irrelevant sentence but also provides diverse perspectives of how a diverse pool of potential readers (i.e., crowd workers) may interpret the writing. For example, the weighted annotations help people understand the gap between their intentions and readers' interpretations. Participant P7 reported that StructFeed helped her realize that the example she used in her essay might confuse other people and she said she would choose a more suitable example in her further revision.

Furthermore, it is clear that StructFeed is more flexible than ML-based method because it can be applied to support the writing of different topics and genres without the needs of training, the availability of corpus or prior knowledge for composing decision rules.

Expert feedback performed worse than crowd feedback?

It is indeed a surprising result. While we don't have hard proofs yet, here are some conjectures from our observations.

- Each revision is limited to 30 minutes. Under the pressure of time and overwhelmed by a large number of editing/comments from the expert, the ESL writers could take

the easy way out by simply clicking to accept the suggested editing without making revisions on their own.

- Expert feedback contains both global issues such as lack of structure or coherence and local issues such as grammatical fix, word choice, etc. The ESL writers tended to focus on the easier fixes of local issues, rather than making the more difficult ones (global issues).
- There might exist the knowledge or communication gap between an expert reviewer and a novice writer. Some participants reported that they want to communicate with experts for further clarifications.

To solve this issue, we are now exploring how to present or filter the comments for facilitating better revision in our ongoing project.

Macro-task vs. micro-task

The quality of feedback generated by leveraging micro-tasks (StructFeed) is shown to be more stable than feedback generated with macro-tasks (expert feedback and free-form crowd worker feedback). For the general crowd feedback condition, a single crowd worker was recruited to generate writing feedback. The quality of feedback cannot be guaranteed to be useful based on varying ability of workers. In contrast, feedback generated by StructFeed aggregated the answers from multiple workers, reducing the risk and uncertainty of obtaining low-quality results.

Conclusion

In this paper, we present StructFeed, a system that helps ESL writers to improve the quality of writings by receiving structural feedback. In the system, a crowdsourcing workflow was proposed to guide crowd workers to annotate topic sentences and relevant keywords through micro-tasks. By aggregating crowd annotations, the system can generate writing hints for directing people to address the structure issues effectively. In a field deployment study, we showed that our system could help ESL writers improve their writings. In addition, people who received feedback from StructFeed outperformed people who received feedback from a expert or a crowd worker. StructFeed enables new kind of writing feedback that cannot obtain from other sources. The work pioneers the design space of generating writing feedback with crowdsourcing mechanisms for ESL writers.

Acknowledgements

This research was supported in part by the Ministry of Science and Technology of Taiwan (MOST 106-2633-E-002-001), National Taiwan University (NTU-106R104045), Intel Corporation, and Delta Electronics, and Advantech.

References

Bernstein, M. S.; Little, G.; Miller, R. C.; Hartmann, B.; Ackerman, M. S.; Karger, D. R.; Crowell, D.; and Panovich, K. 2010. Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*.

Burstein, J.; Chodorow, M.; and Leacock, C. 2004. Automated essay evaluation: The criterion online writing service. *AI Magazine* 25(3):27–36.

Dikli, S. 2006. An overview of automated scoring of essays. *The Journal of Technology, Learning, and Assessment* 5(1).

Dow, S.; Kulkarni, A.; Klemmer, S.; and Hartmann, B. 2012. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work and Social Computing*.

Kaplan, R. B. 1966. Cultural thought patterns in intercultural education. *Language Learning*.

Kim, J., and Monroy-Hernandez, A. 2016. Storia: Summarizing social media content based on narrative theory using crowdsourcing. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*.

Kim, J.; Cheng, J.; and Bernstein, M. S. 2014. Ensemble: Exploring complementary strengths of leaders and crowds in creative collaboration. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing*.

Kulkarni, C. E.; Bernstein, M. S.; and Klemmer, S. R. 2015. Peerstudio: Rapid peer feedback emphasizes revision and improves performance. In *Proceedings of the Second ACM Conference on Learning @ Scale*, L@S '15, 75–84. New York, NY, USA: ACM.

Luther, K.; Tolentino, J.-L.; Wu, W.; Pavel, A.; Bailey, B. P.; Agrawala, M.; Hartmann, B.; and Dow, S. P. 2015. Structuring, aggregating, and evaluating crowdsourced design critique. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work and Social Computing*.

Oshima, A., and Hogue, A. 2013. *Longman Academic Writing Series 4: Essays (5th Edition)*. Pearson Education ESL, 5th edition edition.

Sadler, D. R. 1989. Formative assessment and the design of instructional systems. *Instructional Science* 18(2):119–144.

Teevan, J.; Iqbal, S. T.; and von Veh, C. 2016. Supporting collaborative writing with microtasks. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*.

Tinapple, D.; Olson, L.; and Sadauskas, J. 2013. Critviz: Web-based software supporting peer critique in large creative classrooms. *Bulletin of the IEEE Technical Committee on Learning Technology* 15(1):29–35.

Xu, A.; Huang, S.-W.; and Bailey, B. 2014. Voyant: Generating structured feedback on visual designs using a crowd of non-experts. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing*.