# Analyzing the Revision Logs of a Japanese Newspaper for Article Quality Assessment

**Hideaki Tamori** [1] **Yuta Hitomi** [1] **Naoaki Okazaki** [2] **Kentaro Inui** [3]
[1] Media Lab, The Asahi Shimbun Company
[2] Tokyo Institute of Technology
[3] Tohoku University
{tamori-h, hitomi-y1}@asahi.com, okazaki@chokkan.org
inui@ecei.tohoku.ac.jp

## Abstract

We address the issue of the quality of journalism and analyze daily article revision logs from a Japanese newspaper company. The revision logs contain data that can help reveal the requirements of quality journalism such as the types and number of edit operations and aspects commonly focused in revision. This study also discusses potential applications such as quality assessment and automatic article revision as our future research directions.

## 1 Introduction

Quality journalism deserves serious consideration, particularly given the disruptions of existing publishing companies and the emergence of new publishing companies, citizen journalism, and automated journalism. Although no consensus exists for the definition of quality journalism, Meyer (2009) describes several aspects that constitute quality journalism; for example, credibility, influence, accuracy, and readability. To the best of our knowledge, this is the first attempt to analyze the large-scale revision logs of professionals in the field of journalism. In this study, we explore aspects of quality journalism through analyses of the newspaper article revision logs. More specifically, we analyze the revision processes as editors refine the drafts written by reporters so that they are of publication quality.

While our attempt is still in the early stages, this paper reports the statistics of the actual revisions made by professionals and shows the usefulness of the revision logs. We also discuss the future directions of this research, for example, the potential to present feedback to reporters, extract guidelines for 'good' articles, and develop systems for automatic revision and sentence merging and spitting.

## 2 Analysis of revision logs

This section describes the daily activities of a newspaper company needed to publish articles and the analysis of the revision logs.

### 2.1 Flow of editing and publishing articles

A reporter drafts an article and sends it to an editor, who has over ten years' experience as a journalist. The editor proofreads the article and forwards it to a reviewers section. The reviewers in this section fact-check the article. Finally, designers adjust the article so that they fit in the newspaper and website. In this way, a newspaper article is revised many times from the original submission.

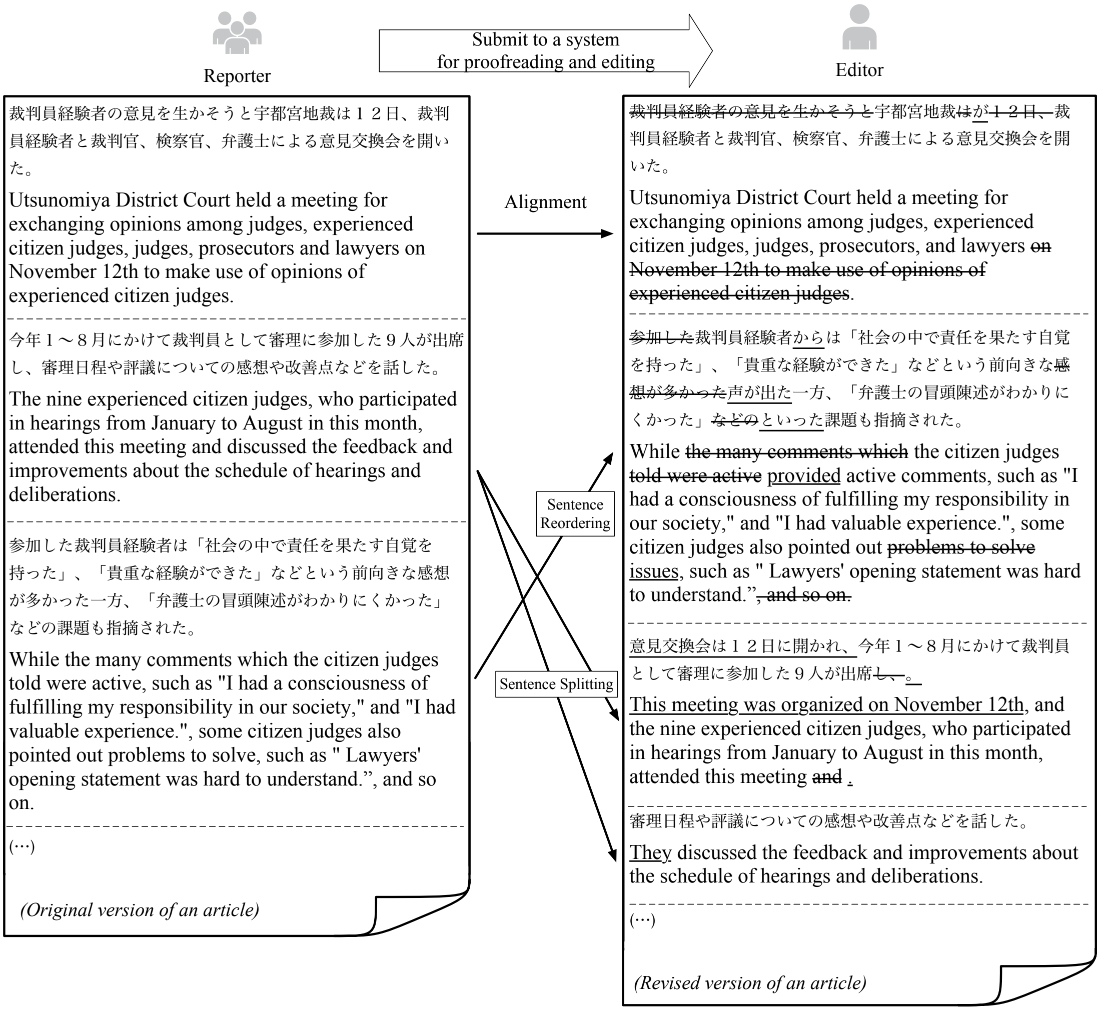
Figure 1 compares the text from an article written by a reporter and the same text after it has been revised by an editor. The editor revises the text using insertion, deletion, and replacement. This example also shows the operations of sentence reordering and splitting.

### 2.2 Aligning original sentences with revised sentences

Revision logs present two versions of an article: the one written by a reporter (the original version) and the final version revised by an editor (the revised version). However, these logs do not provide details about the editing process used to transform the original version into the final version (e.g., individual operations of word insertion/deletion, sentence reordering). Hence, we estimate sentence alignments between the original and revised versions using the *maximum alignment* method (Kajiwara and Komachi, 2016).

The accuracy, precision, recall, and F1 score were 0.995, 0.996, 0.951, and 0.957, respectively, on a dataset consisting of 50 articles in which the correct alignments were assigned by a human.[1]

[1] We chose 0.06 for word similarity threshold and 0.70 for



裁判員経験者の意見を生かそうと宇都宮地裁は１２日、裁判員経験者と裁判官、検察官、弁護士による意見交換会を開いた。

Utsunomiya District Court held a meeting for exchanging opinions among judges, experienced citizen judges, judges, prosecutors and lawyers on November 12th to make use of opinions of experienced citizen judges.

今年１～８月にかけて裁判員として審理に参加した９人が出席し、審理日程や評議についての感想や改善点などを話した。

The nine experienced citizen judges, who participated in hearings from January to August in this month, attended this meeting and discussed the feedback and improvements about the schedule of hearings and deliberations.

参加した裁判員経験者は「社会の中で責任を果たす自覚を持った」、「貴重な経験ができた」などという前向きな感想が多かった一方、「弁護士の冒頭陳述がわかりにくかった」などの課題も指摘された。

While the many comments which the citizen judges told were active, such as "I had a consciousness of fulfilling my responsibility in our society," and "I had valuable experience.", some citizen judges also pointed out problems to solve, such as " Lawyers' opening statement was hard to understand.”, and so on.

(···)

*(Original version of an article)*

~~裁判員経験者の意見を生かそうと~~宇都宮地裁~~は~~が~~１２日、~~裁判員経験者と裁判官、検察官、弁護士による意見交換会を開いた。

Utsunomiya District Court held a meeting for exchanging opinions among judges, experienced citizen judges, judges, prosecutors, and lawyers ~~on November 12th to make use of opinions of experienced citizen judges~~.

~~参加した~~裁判員経験者からは「社会の中で責任を果たす自覚を持った」、「貴重な経験ができた」などという前向きな~~感想が多かった~~声が出た一方、「弁護士の冒頭陳述がわかりにくかった」~~などの~~といった課題も指摘された。

While ~~the many comments which~~ the citizen judges ~~told were active~~ provided active comments, such as "I had a consciousness of fulfilling my responsibility in our society," and "I had valuable experience.", some citizen judges also pointed out ~~problems to solve~~ issues, such as " Lawyers' opening statement was hard to understand.”~~, and so on.~~

意見交換会は１２日に開かれ、今年１～８月にかけて裁判員として審理に参加した９人が出席~~し、~~。

This meeting was organized on November 12th, and the nine experienced citizen judges, who participated in hearings from January to August in this month, attended this meeting ~~and~~ .

審理日程や評議についての感想や改善点などを話した。

They discussed the feedback and improvements about the schedule of hearings and deliberations.

(···)

*(Revised version of an article)*

Figure 1: Comparison of the original and revised versions of some text. In the revised version, the strikethrough and underlined parts indicate deletions and insertions, respectively.

The precision, recall, and F1 score calculated from only the sentence pairs that are changed during revision were 0.926, 0.895, and 0.910, respectively. There may be some room for improving the performance of the alignments but it is sufficient for this analysis.

### 2.3 Data analysis of the revision logs

To analyze the details of the revision processes, we inspected articles published from October 1, 2015 to September 30, 2016. We applied MeCab (Kudo et al., 2004), a Japanese morphological analyzer, with the enhanced dictionary NEologd[2], to split the sentences into words (morphemes) and recognize their parts-of-speech.

The dataset analyzed in this study contains 120,331 articles with 1,903,645 original sentences and 1,884,987 revised sentences. The dataset consists of a Japanese newspaper's articles[3], which have a mixed domain (genre) of the news, and most of the articles have the same writing style. We obtained 2,197,739 sentence pairs using the alignment method described in Section 2.2. The number of aligned pairs is larger than that of the sentences because an original sentence can be aligned to multiple sentences in the revised version. About half of the sentence pairs (1,108,750) were unchanged during the revision process, and the remaining pairs (1,088,989) were changed. In this section, we report the statistics of the edit operations in the changed pairs. We found that newspaper companies produce a huge number of sentences, about half of which are revised, for analy-

sentence similarity threshold, optimizing the F1 score on a development set consisting of 150 articles. We used word embeddings that are pre-trained by the original articles and revised articles to compute sentence similarity.

[2] https://github.com/neologd/mecab-ipadic-neologd

[3] The Asahi Shimbun Company provided this dataset.

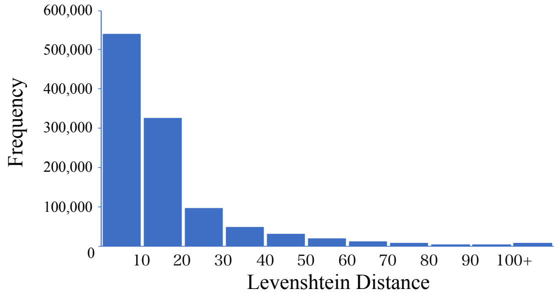


Figure 2: Distribution of the Levenshtein distance of changed sentence pairs.

Original sentence:
市場に同じ魚が出回りすぎると、魚の単価が下がってしまう。
If the same kind of fish is distributed in large quantities
in the market, the unit price of the fish manages to decrease.

Revised sentence:
~~市場に~~同じ魚~~が出回りすぎる~~の水揚げが重なると、魚の単価が
下が~~ってしまう~~る。
If the same kind of fish is ~~distributed~~ landed in large quantities
~~in the market~~, the unit price of the fish ~~manages to decrease~~
is decreasing.

Figure 3: An example of the original and revised sentence pair whose Levenshtein distance is 15.

sis just within a year.

Figure 2 presents the distribution of the Levenshtein distance between the original and revised sentences. The mean of the Levenshtein distances of the revised pairs (15.04) indicates that the dataset includes many examples in which drafts are deeply edited by the editors. Figure 3 is an example of the sentence pair which has the mean of the Levenshtein distance of the dataset (15).

Table 1 lists the number of insertions, deletions, and replacements, according to the number of words involved in the edit operations. We found that 56.20% of the total edit operations were replacements for one or two words, and this fact indicates that editors revised these articles with impressive attention to detail.

Table 2 shows the number of edit operations separated by different part-of-speech. The most frequent target for revisions is nouns, followed by particles (postpositions). This result indicates that revisions in terms of both content and readability are important for improving the quality of articles.

| # of words | Insertion | Deletion | Replacement |
|---|---|---|---|
| 1 | 139,790 | 160,975 | 1,424,118 |
| 2 | 118,261 | 151,641 | 303,293 |
| 3 | 57,397 | 53,789 | 115,525 |
| 4 | 35,272 | 31,719 | 75,909 |
| 5 | 21,339 | 20,435 | 33,805 |
| 6 | 13,295 | 14,756 | 21,419 |
| 7 | 14,599 | 13,030 | 24,400 |
| 8 | 8,631 | 9,301 | 10,707 |
| 9 | 10,196 | 10,760 | 8,475 |
| Over 10 | 48,754 | 60,523 | 61,387 |
| **Total** | 467,534 | 526,929 | 2,079,038 |

Table 1: Number of edit operations with respect to the number of words involved.

| Tag | Count |
|---|---|
| Noun | 1,255,113 |
| Noun + Noun | 174,306 |
| Particle | 157,840 |
| Symbol | 128,584 |
| Noun + Particle | 106,548 |
| Verb | 85,709 |
| Symbol + Noun | 47,635 |
| Particle + Noun | 42,714 |
| Particle + Verb | 41,342 |
| Prefix | 41,194 |
| Noun + Symbol | 37,580 |
| Verb + Auxiliary | 20,836 |
| Auxiliary | 18,145 |
| Noun + Verb | 14,153 |
| Adverb | 9,009 |
| Others | 101,714 |

Table 2: Distribution of parts-of-speech as targets for the edit operations involving one or two words.

## 3 Future directions for quality assessment and automatic article revision

There are several possible future directions for the utilization of the revision logs.

### 3.1 Feedbacks to reporters

We can use the revision logs for improving the writing skills of reporters. An interesting finding in the revision logs is that the articles of young reporters (1–3 years' experience) tend to be revised more than those of experienced reporters (31–33 years' experience): the mean Levenshtein distances of these young and experienced reporters are 15.82 and 12.95, respectively. As exemplified by this finding, the revision logs can indicate the main types of revisions that a particular group of reporters or an individual reporter receives. We will explore the potential of the revision logs for assessing the writing quality of a reporter and presenting them with feedback.

### 3.2 Establishing guidelines for writing articles

Most textbooks on Japanese writing (including the internal handbook for reporters produced by the newspaper company) recommend that a Japanese sentence should be 40 to 50 characters long (Ishioka and Kameda, 2006). We could confirm that the newspaper articles satisfy this criterion: the revised sentences are 41.10 characters long on average. In this way, we can analyze the revision logs to extract various criteria for establishing the guidelines for 'good' articles.

### 3.3 Automatic article revision within sentences

Another future direction is to build a corpus for improving the quality of articles. The revision logs collected for a year (excluding duplicates) provide 517,545 instances of replace operations, 79,639 instances of insertions, and 54,111 instances of deletions that involve one or two words. Table 3 shows some instances of the replace operations. It may not be straightforward to use the revision logs for error correction because some edit operations add new information and remove useless information. Nevertheless, the logs record the daily activities of how drafts are improved by the editors. In future, we plan to build an editing system that detects errors and suggests wording while the reporters write drafts. We can use natural language processing techniques for these tasks because local error correction has been previously researched (Cucerzan and Brill, 2004).

| **Original** | **Revised** |
|---|---|
| 同政府関係者<br>this Government officials | 韓国政府関係者<br>Korean Government officials: specification |
| 放射線汚染<br>contamination by radial ray | 放射能汚染<br>radiologically contamination |
| 破顔し<br>broke into a smile | 笑顔で話し<br>spoke with a smile: simplification |
| バラティ<br>Varety | バラエティー<br>Variety: typo |
| タンパク質<br>Protain: written in Katakana and Kanji | たんぱく質<br>Protain: written in Hiragana and Kanji |
| 買えた<br>could buy | 買える<br>can buy |

Table 3: Examples of commonly replaced words/phrases.

### 3.4 Automatic sentence merging and splitting

The alignment method found 69,891 instances of sentence splitting (wherein an original sentence is split into multiple sentences) and 68,550 instances of sentence merging (wherein multiple original sentences are merged into one sentence). Table 4 shows examples of sentence splitting and merging. We observe some sentences are also compressed during sentence merging and splitting. We can use these instances as a training data for building a model for sentence splitting and merging (with compression), and this may be an interesting task in the field of natural language processing.

## 4 Conclusion

In this paper, we explored the potential of the revision logs of a newspaper company for assessing the quality of articles. In addition to presenting the revision logs statistics, we discussed the future directions of this work, which include feedback to reporters, guidelines for 'good' articles, automatic article revision, and automatic sentence merging and splitting.

## References

Silviu Cucerzan and Eric Brill. 2004. Spelling correction as an iterative process that exploits the collective knowledge of web users. In *Proc of EMNLP*. pages 293–300.

Tsunenori Ishioka and Masayuki Kameda. 2006. Automated Japanese essay scoring system based on articles written by experts. In *Proc of ACL*. pages 233–240.

Tomoyuki Kajiwara and Mamoru Komachi. 2016. Building a monolingual parallel corpus for text simplification using sentence similarity based on alignment between word embeddings. In *Proc of COLING*. pages 1147–1158.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proc of EMNLP*. pages 230–237.

Philip Meyer. 2009. *The vanishing newspaper : saving journalism in the information age*. University of Missouri Press, Columbia.

| | |
|---|---|
| Splitting | |
| (S1) | 新たな窓口を設けるなど内部通報制度も強化し、通報は問題が発覚する前の 88 件 (14 年度) から 263 件 (15 年度) と 3 倍に増えた。<br>They enhance whistle-blowing systems by providing such as new counseling offices, and the number of whistle-blowing was increased three times from 88 in 2014, in which this issue was found out, to 263 in 2015. |
| (S2) | 新たな窓口を設けるなど内部通報制度も強化。通報は 2015 年度に 263 件と、不正会計問題の発覚前の 14 年度の 88 件から約 3 倍に増えたという。<br>They enhance whistle-blowing systems by providing such as new counseling offices. As a result, the number of whistle-blowing was increased three times from 88 in 2014, in which this issue was found out, to 263 in 2015. |
| Merging | |
| (M1) | 同署によると、事務所南側 1 階の窓が割られ、室内にある防犯カメラのモニター 4 台がすべて壊されていた。食器棚も倒され、食器が散乱していたという。<br>Police said that the window on the first floor of the office south is broken, and all four displays for the security camera was destroyed. Police also said that the cupboard was knocked down, and the dished are scattered in the room. |
| (M2) | 署によると、室内の防犯カメラのモニター全 4 台が壊され、食器棚が倒れて食器が散乱していた。<br>Police said the all four displays for the security camera was destroyed, the cupboard was knocked down, and the dished are scattered in the room. |

Table 4: Examples of sentence splitting and merging. Sentences S1 and M1 are the original sentences, and S2 and M2 are the revised sentences. In the merging example, we can also observe the sentence compressing; the part "the window on the first floor of the office south is broken" was eliminated in M2.