Evaluation of Effectiveness of Majority Voting Method in Crowdsourcing-based Subtitling Methods

Kantaro Mera\textsuperscript{1a)}, Ryuya Itano\textsuperscript{2}, and Takahiro Koita\textsuperscript{1}

\textsuperscript{1}Graduate School of Engineering, Doshisha University
1-3 Tatara Miyakodani, Kyotanabe, Kyoto 610-0394, Japan

\textsuperscript{2}Faculty of Science and Engineering, Doshisha University
1-3 Tatara Miyakodani, Kyotanabe, Kyoto 610-0394, Japan

\textsuperscript{a}) ctwf0123@mail4.doshisha.ac.jp

Abstract: Crowdsourcing is an effective way to generate subtitles for videos with multiple speakers. This study proposes and evaluates a word-by-word majority voting method as a new approach to creating high-accuracy subtitles from multiple types of subtitles created by multiple workers in crowdsourcing. In the word-by-word majority voting method, the accuracies of the subtitles themselves remarkably vary between high and low, since even a slight difference in word order, for example, can significantly reduce incorrect generation. This remarkable difference shows that the proposed method has potential to generate appropriate subtitles for videos having multiple speakers.

Keywords: crowdsourcing, crowd work, captions, language barrier

Classification: Multimedia Systems for Communications

References

[1] Kantaro Mera, Taichi Sugimura, and Takahiro Koita, A proposed system for machine translation by crowdsourcing, Proc. of the Networking, Computing, Systems, and Software, Vol. 10, No. 1, pp. 25–26, 2021.

[2] Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matthew Lease, and John J. Horton, The Future of Crowd Work, Proc. of the Conference on Computer Supported Cooperative Work, pp. 1301–1318, 2013.

[3] Vikas C. Raykar and Shipeng Yu, Ranking Annotators for Crowdsourced Labeling Tasks, Advances in Neural Information Processing Systems, Proc. of the Conference on Neural Information Processing Systems, pp. 1809–1817, 2011.

[4] Lili Wang, Ye Yang, and Yong Wang, Do Higher Incentives Lead to Better Performance? - An Exploratory Study on Software Crowdsourcing, Proc. of the ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, pp. 1–11, 2019.
1 Introduction

Subtitles are needed to overcome language barriers in video viewing. Existing automatic subtitling systems have been implemented on such platforms as YouTube where videos are uploaded and subtitles are created from a video’s audio data. Although existing systems can correctly convert audio data from a single speaker into subtitles, they cannot correctly convert audio data from multiple speakers into subtitles. The existing systems cannot generate subtitles by decomposing the audio data synthesized from multiple sounds produced by multiple speakers. Therefore, we previously proposed a method using crowdsourcing that can incorporate skilled listeners in its system’s operation [1]. In this paper, this method is called the Crowdsourcing Subtitles System (CSS). Crowdsourcing, which requests work from an unspecified number of individuals through the internet, is widely used in such fields as the collection of teaching data for machine learning [2]. In a previous study, the accuracy of multiple subtitles created with CSS was compared to those created with existing systems such as YouTube, and CSS was shown to create more accurate subtitles [1]. However, CSS is not able to create a single appropriate subtitle from the multiple subtitles created by the crowd workers. To build a subtitle creation system using CSS, a single high-precision subtitle must be chosen from multiple subtitles created by crowdsourcing. In this paper, we propose a word-by-word majority voting method as a way to produce a single subtitle from multiple types of subtitles created by CSS, and we also show its effectiveness. The majority voting method is one of the most useful data-quality methods in crowdsourcing, where multiple responses are solicited per datum and a majority vote is taken [3].

2 Previous study

In our previous study [1], CSS was shown to be sometimes more effective for creating subtitles than existing automatic subtitling systems, such as YouTube. CSS consists of the following three steps:

1. Divide a video into segments of a certain duration.
2. Perform a transcription task on the divided video segments by crowdsourcing.
3. Divide the created subtitles into words, number the words starting from the first one, compare the words of the same-order number, and score them if they correctly match.

Twenty-three workers crowdsourced the creation of subtitles for six videos. Then, the average accuracy of the subtitles created by the crowdsourced workers was compared to the accuracy of the subtitles created by YouTube’s automatic subtitle-creation system. Consequently, the accuracy of the subtitles created by the crowdsourced workers was found to be at least three times higher than that of the subtitles created by YouTube.

Our previous study showed that CSS can effectively create subtitles. To develop a subtitling system using CSS, multiple subtitles produced by crowdsourcing need to be evaluated toward selecting a single one. In this study, as an evaluation scheme, the majority voting method is adopted to efficiently select the appropriate subtitle from multiple responses. In addition, video script is necessary for calculating the
accuracy but video script is not in the actual subtitling stage. So, when creating subtitles, it is necessary to compare accuracy as words, not as numbers. Therefore, as an evaluation method, the majority voting method is adopted to make a word-by-word comparison.

3 Experiment

Fifty-five tasks were requested from Amazon Mechanical Turk, with ten people assigned to each task. Here, the task was to transcribe a three-minute video. Amazon Mechanical Turk was launched by Amazon Web Services in 2005, and now Turk is the world’s largest crowdsourcing platform. The number of tasks is the number required to show significance at a confidence level of 80% and a tolerance of 5%, based on the assumption that 500 hours of video are uploaded to YouTube every minute. The video used has four speakers. For verification, a version of the video having correct subtitles was used, and instead of actually performing the majority voting method, the correct subtitles and the crowdsourced subtitles were compared word by word. The agreement rate was derived from the number of correct answers, which revealed that the subtitle with the highest agreement rate matched the one chosen by the majority voting method. The following gives the specific procedure of the experiment:

(1) Request crowdsourced workers to create subtitles.
(2) Compare the created subtitles with the correct ones using word units.
(3) The subtitle with the highest rate of agreement with the correct answer is considered the final product.

![Accuracy Comparison](image.png)
This section explains how the matching rate is calculated in this experiment. The agreement rate is calculated as

$$ P \div W \times 100 = MR. \quad (1) $$

The original subtitle prepared for the matching rate comparison is compared with the created subtitle, starting from the first word in each. If the two same-order words are identical, one point is added; if they are different, no point is added. The final result is the ratio of the number of obtained points (P) to the total number of words (W), which forms the subtitle matching rate (MR).

**4 Results**

The comparison of matching rates in ten tasks is shown in Fig.3(a), where the performance in each task by the proposed system is compared with those of existing systems (YouTube and Amazon Transcribe). The proposed method, i.e., CSS and the word-by-word majority voting method, showed the greatest number
of tasks with the most accurate subtitles among the three methods. Some of the subject videos were spoken by a single speaker for most of the three minutes, and in such cases, the subtitles created by a method using the two existing systems recorded higher accuracy. However, the proposed system recorded about three-times higher accuracy than the existing system for videos that contain many scenes where multiple speakers are having a discussion.

Therefore, when focusing solely on accuracy, the proposed method cannot be said to have numerically high accuracy. This is due to the method of calculating the matching rate based on whether the words in the correct and created subtitles match sequentially from the first word. In this method, if the third word of the created subtitle is the same as not only the third but also the second or fourth word of the correct subtitle (i.e., one before/one after), the subtitle is considered correct. With this method of calculation, word numbers may shift significantly due to contractions or other writing styles, and all words outside the corresponding number range would be incorrect, resulting in a significant loss of statistical accuracy. Accordingly, the matching rate might be improved by increasing the range of the order of the words to be compared to two or three, rather than just one, before or after each word.

The subtitles created by the proposed system, YouTube, and Amazon Transcribe are summarized for the factors of agreement, cost, time, and speaker identification (Figure 3(b)). The proposed system creates subtitles at a cost of $1.0 for 10 people for a 3-minute video, while YouTube and Amazon Transcribe create them for almost free. However, as shown in previous research [4], increasing rewards or incentives would not improve quality. Rather, the cost could be reduced by effectively filtering the workers. On the other hand, beyond filtering, if the worker was a professional transcriptionist, the cost would be considerably higher than the proposed system. Thus, CSS is more cost-effective compared to other existing systems. As for the time required for creation, the proposed system takes about 24 hours, while the existing systems take a few minutes. However, in the case of YouTube, not all videos have subtitles. Furthermore, the proposed system is the only one that can create subtitles that identify which speaker says which words (speaker identification).

5 Conclusion

In this study, by applying CSS to creating subtitles for multi-speaker videos, we show that the word-by-word majority voting method is effective for selecting appropriate subtitles from multiple subtitles. The results of the analysis in terms of accuracy, cost, creation time, and speaker identification show that the proposed method is more useful than YouTube, Amazon Transcribe, and professional translators. The challenge for the future is accuracy per cost, or cost-effectiveness. Therefore, it is necessary to focus on how to issue tasks and select crowdsourcing workers in order to create more accurate subtitles.
(a) Comparison of agreement rates among methods (partial)

(b) Multidimensional analysis of methods

|                           | Proposed System | YouTube | Amazon Transcribe |
|---------------------------|-----------------|---------|-------------------|
| Matching Rate (%)        | 35.3            | 20.4    | 18.5              |
| Cost ($)                 | 1.0             | 0       | 0                 |
| Run Time (h)             | 24              | 0       | 0                 |
| Speaker Identification   | ✔               | Not Applicable | Not Applicable |

Fig. 3 Results analysis