Progressive Language Model Adaptation for Disaster Broadcasting with Closed-captions

Takahiro Oku, Yuya Fujita, Akio Kobayashi and Shoei Sato
NHK (Nippon Hoso Kyokai; Japan Broadcasting Corporation.)
Science and Technology Research Laboratories, Setagaya 157-8510, Tokyo, Japan
E-mail: {oku.t-le, fujita.y-gc, kobayashi.a-fs, satou.s-gu}@nhk.or.jp Tel: +81-3-5494-3369

Abstract— **This paper describes a progressive language model (LM) adaptation method for transcribing broadcast news in a sudden event such as a massive earthquake. In a practical automatic speech recognition (ASR) system, the new event whose linguistic contexts are not covered with the LM often causes a serious degradation of the performance. Furthermore, there might be not enough quantities of training texts for conventional LM adaptation such as linear interpolation. Then, we propose a new LM adaptation method by using ASR transcriptions as unsupervised training texts in addition to the online manuscripts written by reporters. The proposed method employs a progressive update procedure, which adapts LMs in an unsupervised manner by using every set of transcriptions in a short period for the purpose of immediate use of the adapted model. The method also uses the online manuscripts in order to adapt the LM and add new words into the vocabulary. Experimental results showed that the proposed progressive LM adaptation method reduced relatively a word error rate by 8.2% compared with the conventional LM adaptation method with the online manuscripts only.**

I. INTRODUCTION

For Japanese broadcasters including NHK (Japan Broadcasting Corporation), real-time closed-captioning of live broadcast programs is still a challenging issue of information accessibility for the hearing-impaired and elderly people. In particular, during disasters, it is a powerful means of conveying information. In response to such social demands, NHK has been developed real-time closed captioning systems using automatic speech recognition (ASR).

On March 11, 2011, a massive earthquake hit the eastern Japan and caused both an extreme tsunami disaster and a terrible nuclear power plant disaster. In the face of such heavy disasters, all Japanese broadcasters passed the news along for days and nights. To keep a steady closed captioning service for live news programs, NHK operated a prototype of the current closed-captioning system and provided the captions for some programs.

The current system launched on March 14, 2012 is being operated for daily straight news programs. It is connected to NHK's internal news network and automatically assembles the latest online manuscripts written by reporters. To achieve higher ASR performance, language models (LMs) for every news programs are adapted with the online manuscripts. Generally, similar topics extends over several days in typical news shows, and thus, LM adaptation such as linear interpolation and count-merge methods using the latest manuscripts has been proved to be effective for reducing word error rates [1]. In the disaster broadcasting, meanwhile, we may not obtain enough quantities of the online manuscripts to maintain the ASR performance because the disaster itself is an unprecedented event. A widely referred online manuscript must go through several modification or confirmation processes before its release, while the extent of damage is gradually revealed. This leads to an inconsistency between the contents of released online manuscripts and the changing situation. Thus, ASR would have to deal with such an emergency event which is not covered with the LM, and the ASR consequently suffers serious performance degradation. Although an LM adaptation method such as linear interpolation is employed to avoid the degradation caused by such a changing situation, there might be not enough quantities of training texts for LM adaptation. Therefore, not only the online manuscripts but also ASR outputs should be used for LM adaptation. According to the past researches (*e.g.* [2][3]), conventional linear-interpolated LM adaptation using ASR outputs has yielded a significant word error reduction. However, there has been less interest in LM adaptation from a diachronic perspective of the event spanning hours to days such as a disaster.

In this paper, we propose a new LM adaptation method for transcribing news programs such as disaster broadcasting. The proposed method employs a progressive update procedure, which successively adapts LMs in an unsupervised manner by using every set of ASR outputs obtained in the most recent short period. The method also uses the online manuscripts to achieve a high ASR performance as well as to add new words into the vocabulary.

The rest of this paper is organized as follows. In Section II, we describe an overview of the ASR system. In Section III, an LM adaptation method for disaster broadcasting is proposed. In Section IV, we investigate the transition speeds of a situation in order to examine whether LMs should be constantly adapted in a short period. In Section V, the ASR performance of the progressive LM adaptation method was evaluated by using recognition results as unsupervised training data. In Section VI, our prototype device to realize

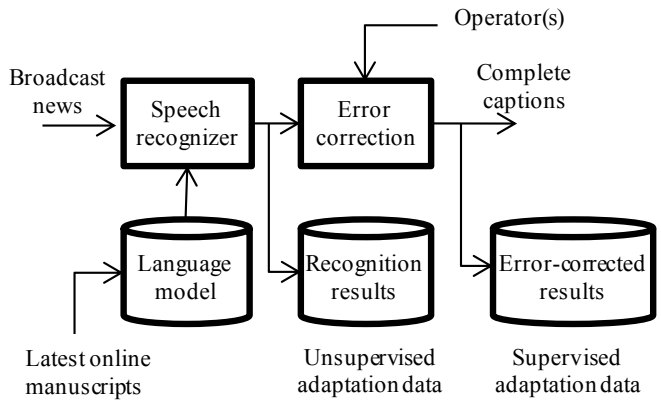

Fig. 1 Closed captioning system

the proposed method was introduced. Section VII concludes the study, and future works are mentioned.

II. CLOSED CAPTIONING SYSTEM

In this section, we introduce an overview of the ASR system for closed captioning. The system, installed in our broadcasting station, is currently operated for daily news programs. It generates accurate text for the captioning from the utterances spoken during the programs. As illustrated in Fig. 1, the speech recognizer is tuned for the daily news by LM adaptation. The word error rate of the recognizer is less than 5%. Misrecognized words are immediately corrected by operators. Typically, an operator corrects an error word every 6 seconds by using a smart error correction interface as described in Section II.B. Then, the error free caption data is transmitted to viewers.

A. Speech Recognizer

Captions of live broadcasts, for which prepared text is not able to transmit due to the existence of spontaneous speech such as conversations, are displayed with a delay of around ten seconds on screens of viewers. The speech recognizer progressively outputs the latest available words so as to shorten the undesirable delay. The progressive recognizer described in [4] implements a low latency text output from a speech input while maintaining a high performance. To reduce the delay caused by the operations of the following manual error correction, accuracy of transcriptions is required to be above 95% by using domain adapted LMs. The LMs for all news programs are adapted to the latest online manuscripts. In our case, online manuscripts written by reporters all over the country are uploaded to the server computer on NHK's internal news network. The captioning system downloads the online manuscripts from the server, and it automatically constructs an LM adapted to them. The LM is adapted by using a weighted mixture of long-term news scripts and latest scripts [1].

B. Manual Error Correction

For efficient manual confirmation and error correction, the error correction system has a touch panel that can be used to identify error words and an ordinary keyboard for inputting words to be substituted. Homonyms of the identified words and histories of the corrected words are also displayed on the touch panel, so as to shorten the operation time needed to substitute such words by just touching the panel. As the system performance strongly depends on a speaking style of TV program to be captioned, the correction system flexibly relies utterances to be corrected among one or two operators [5].

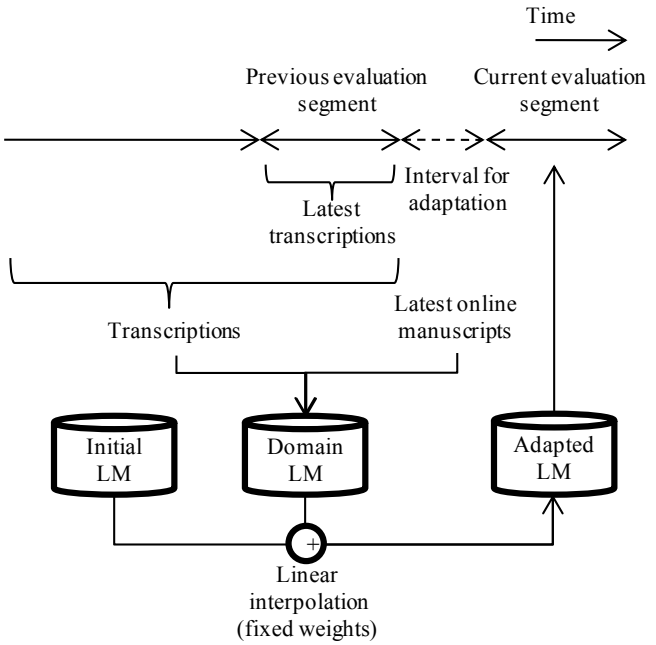

Fig. 2 Progressive LM adaptation

III. PROGRESSIVE LM ADAPTATION FOR DISASTER BROADCASTING

In this section, we describe a new *progressive* LM adaptation method for transcribing news programs in the emergency such as disaster broadcasts. In disaster broadcasting, we may not obtain enough quantities of online manuscripts to keep the performance high. As the widely referred online manuscript has several confirming processes before its release, the situation related to the disaster, such as a spread of damage from the disaster and a disorder of lifeline, tends to be changed before the release of the online manuscripts. The anchor has to announce the current situation mixed with comments written in the online manuscripts. Consequently, the predictive performance of LMs will be degraded seriously due to the existence of new events including unknown linguistic contexts. Thus, LMs should be updated and adapted to follow up the current contexts rapidly and progressively.

To reflect the latest disaster information into the LMs as rapidly as possible, we adapt the LMs progressively by using recognition results or error-corrected transcriptions obtained in a short period. Fig. 2 shows a diagram of progressive LM

Table I 5 topics constituting disaster broadcast

Topics	Content of commentary
Victim	Informing about damage from the disaster
Earthquake	Informing about earthquake intensities
Tsunami	Informing about hitting time of tsunami, etc.
Lifeline	Informing about traffic anomaly or electric power outage
Nuke	Informing about the nuclear accident

adaptation method. In our method, the latest recognition results (or error-corrected results) are obtained in every ten to sixty minutes. N-grams are estimated from the accumulated amount of the transcriptions including the latest ones and online manuscripts in the unsupervised manner. Then, the estimated domain LM is interpolated with a baseline LM, which is obtained from a large amount of online manuscripts and transcriptions, to yield a final adapted LM. Although the mixture weights are fixed in our experiments, they can be estimated from transcriptions according to cross-validation. Since our closed-captioning system can seamlessly switch multiple recognizers that have different configurations, the adapted LM is immediately applied to the latest speech segment [7]. Although it takes time to construct the adapted LM in practice, we set the training time to zero for convenience in our experiments. The adaptation procedure with error-corrected results instead of recognition results, as the domain LM can be estimated in the supervised manner, the ASR performance will be achieved higher. Another advantage for use of error-corrected results is that new words can be added into the vocabulary, which is not covered with the online manuscripts.

IV. Investigation for Disaster Broadcasting

In advance of experiments with our proposed adaptation method, we investigate how LMs affect the linguistic contexts of disaster broadcasts distant from the time the models were adapted. The purpose of this investigation is to examine whether constant LM adaptation should be performed in a short period. This is because the linguistic contexts will be expected to change drastically from the diachronic perspective.

In this paper, we used utterances of anchors that are extracted from the NHK's disaster broadcasting aired from 14:46 to 23:00 on March 11, 2011. We selected utterances of anchors while the other materials such as field reports and interviews were excluded. This is because the anchors' utterances contain crucial information about the disaster, and are also clear enough to generate captions by the ASR. The selected utterances can be categorized according to topics of *victim*, *earthquake*, *tsunami*, *lifeline*, and *nuke*. Brief explanations about the topics are shown in Table I.

We examined transition speeds of situations based on perplexities as measurements. The procedure of calculating perplexities is shown in Fig. 3. They were calculated for evaluation segments determined by a moving window. The

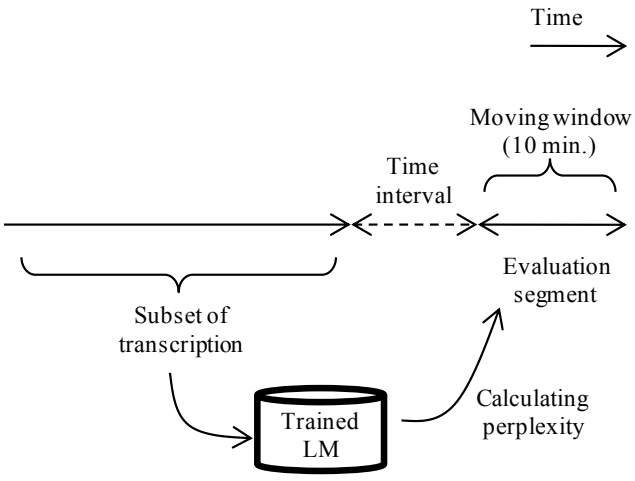

Fig. 3 Procedure of calculating perplexities for examining transition speeds of situations

width of an evaluation segment was 10 minutes and shift width was 5 minutes. To evaluate each segment by perplexity, we trained LMs by using a subset of transcriptions. The perplexities were measured while we varied an interval parameter, a time interval between the beginning of an evaluated segment and the end of an LM training period, where the subset of transcriptions is included. The results are shown in Fig. 4. The perplexities corresponding to the topics, *earthquake*, *tsunami*, *lifeline* are shown in this figure. The horizontal axis indicates the time intervals. As shown in the figure, the perplexities were degraded against the delayed training transcription. The perplexities of *lifeline* were more rapidly increasing than those of *earthquake* and *tsunami*. It indicates that the situation of *lifeline* changed more drastically than other topics. Thus, the mismatch between the LM and the contexts was increased as the time interval was increased. Actually, in the disaster broadcasting, situation constantly changed in *lifeline*. On the other hand, the situation was not

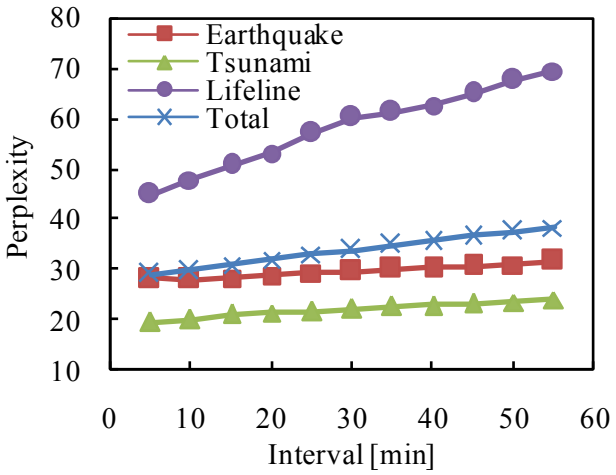

Fig. 4 Changes of perplexities: Perplexities of three topics against the interval. The interval is a time interval between the beginning of an evaluation segment and the end of LM training period.

Table II Specification of evaluated utterances of anchors for 5 topics

Topic	#words	PP	OOV(%)	WER(%)
Victim	20k	53	0.9	8.1
Earthquake	9k	59	0.7	10.7
Tsunami	12k	76	1.0	11.2
Lifeline	10k	75	1.2	10.8
Nuke	3k	53	0.5	7.4
Total	54k	63	0.9	9.8

updated so frequently in *earthquake* and *tsunami*, and the same information tended to be announced several times. These are consistent with the investigated results that the perplexities of *lifeline* were more rapidly increasing than the others.

V. EXPERIMENTS

A. Experimental Setup

We examined the proposed LM adaptation method by ASR evaluation. The acoustic inputs are parameterized into 39 dimensional vectors: 12 mel frequency cepstral coefficients (MFCCs) with log-power and their first- and second-order differentials. Gender dependent acoustic models (AMs) of tri-phone hidden Markov models (HMMs) are used by the recognizer. The recognizer automatically detects the gender of the speaker, allowing for the use of more accurate gender-dependent AMs [7]. The recognizer uses an n-gram LM, and the recognition engine has a 2-pass decoder that uses bigrams and trigrams in the first and second passes, respectively.

The initial trigram LM, which was not adapted by any online manuscripts, was trained on Japanese broadcast news manuscripts and transcriptions (239M words), and the vocabulary size was set to 60k. Table II lists the evaluated utterances of anchors mentioned in the previous section. The perplexities (PP), out of vocabulary (OOV) rates, and word error rates (WER), were measured by the initial trigram LM. The LM adaptation was performed every 10 minutes. The updated LM was obtained by the interpolation between the initial LM and the LM estimated from training texts. The interpolation weight for the initial LM was set to 0.7. According to the composition of online manuscripts, recognition results and error-corrected results, we estimated following three types of LMs:

1. Baseline: LM estimated from online manuscripts only
2. Unsupervised: LM estimated from online manuscripts and recognition results
3. Supervised: LM estimated from online manuscripts and error-corrected results

Because of difficulty to obtain actual online manuscripts at the time of the disaster on March 11, 2011, we used news texts with their update time on the NHK website as simulated data instead of the online manuscripts. At every LM updating, an average of 75 utterances (2.3k words) of online manuscripts was accumulated as the training texts. The OOV rate was reduced by 0.1 % on average using the online manuscripts.

B. Experimental Results

Table III compares the WERs and OOV rates for each topic. In the supervised adaptation, the WER was the most improved for every topic and was lower limit in our proposed method. The unsupervised adaptation reduced WER for each topic compared to the baseline. WERs of baseline and unsupervised adaptation in *Total* were 8.5% and 7.8%, respectively (a word error reduction rate of 8.2%). Obviously, the proposed method reduced WERs significantly.

Next, we divided training texts in unsupervised adaptation into two datasets, online manuscripts and recognition results, and an improvement by using each dataset was examined. The results are shown in Table IV. To ensure equal vocabulary condition between the two datasets, new words obtained from online manuscripts were not registered to the vocabulary. The WER by using recognition results was reduced more than by the online manuscripts for all topics except *nuke*. This result shows that evaluated utterances matched to an n-gram estimated from recognition results more than one estimated from online manuscripts.

Table V shows WERs for several updating intervals. The LM was adapted in the unsupervised manner. The WERs are

Table III LM adaptation results: Comparison of recognition performance between unsupervised and supervised LM adaptations. In addition to the online manuscripts for the baseline, the unsupervised adaptation utilized the recognition result and the supervised utilized the transcriptions of the speech, which are yielded by the error correction

Topics \ LM adaptation	Baseline (online manuscripts)		Unsupervised (online manuscripts & recognition results)		Supervised (online manuscripts & error-corrected results)	
	WER(%)	OOV(%)	WER(%)	OOV(%)	WER(%)	OOV(%)
Victim	7.1	0.7	6.4	0.7	5.8	0.7
Earthquake	10.0	0.6	9.2	0.6	8.7	0.6
Tsunami	8.8	0.8	7.8	0.8	6.6	0.7
Lifeline	9.4	1.1	8.8	1.1	7.1	0.7
Nuke	6.5	0.4	6.8	0.4	6.5	0.4
Total	8.5	0.8	7.8	0.8	6.8	0.7

Table IV LM adaptation results: Comparison of recognition performance on the equal vocabulary condition.

	Online manuscripts		Recognition results	
	WER(%)	OOV(%)	WER(%)	OOV(%)
Victim	7.5	0.9	6.8	0.9
Earthquake	10.1	0.7	9.4	0.7
Tsunami	9.1	1.0	8.2	1.0
Lifeline	9.6	1.2	9.4	1.2
Nuke	7.5	0.5	7.8	0.5
Total	8.8	0.9	8.2	0.9

measured for all topics. They were increased as the updating interval was increased. It indicates that the LM adaptation every 10 or 20 minutes was suitable to maintain ASR performance. In the Great Eastern Japan Earthquake, a variety of information about the disaster was broadcasted all the time. There is no precedent for such a massive disaster, so that it is considered that adaptation every 10 or 20 minutes maintains ASR performance for the other disaster broadcasts.

Table V LM adaptation results against updating intervals

Updating interval (min.)	10	20	30	45	60
WER (%)	7.8	7.8	8.0	8.1	8.4

Fig. 5 shows the temporal changes of the WERs for four conditions: baseline adaptation, unsupervised adaptation every 30 minutes, unsupervised adaptation every 10 minutes, and supervised adaptation every 10 minutes. The WERs of every one hour speech segment were measured. Fig. 6 indicates the word error reduction rates compared with the baseline. During the first 4 hours, the WER was more reduced in unsupervised adaptation than baseline, and adaptation every 10 minutes showed better performance than the 30 minutes. On the contrary, during the next 4 hours, there is little difference in WERs between the baseline and the unsupervised adaptation. For all conditions, ASR performance was degraded around 15 and 19 o'clock. The degradation around 15 o'clock was caused by background noise such as tsunami roar or engine noise of a rescue helicopter.

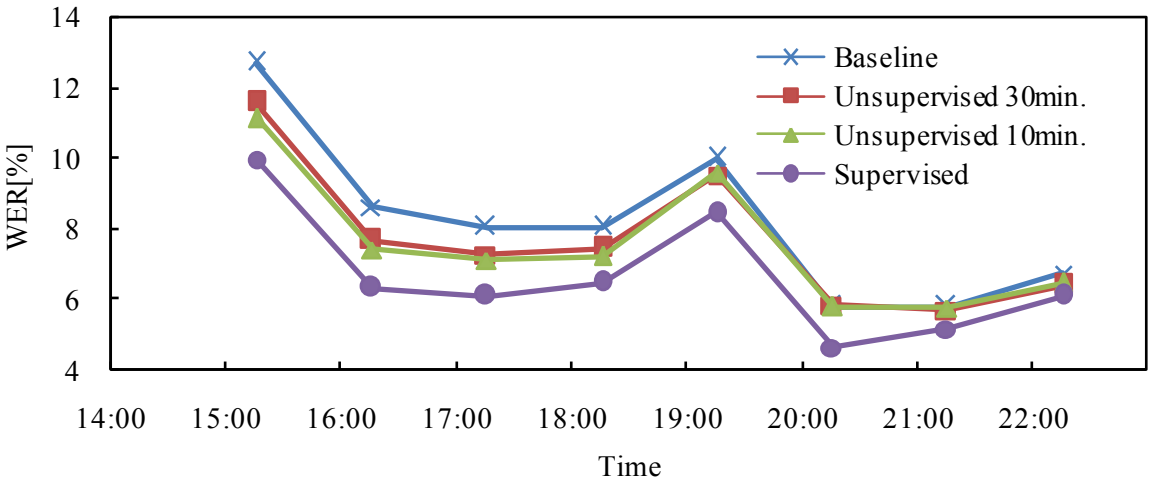

Fig. 5 Changes of WERs

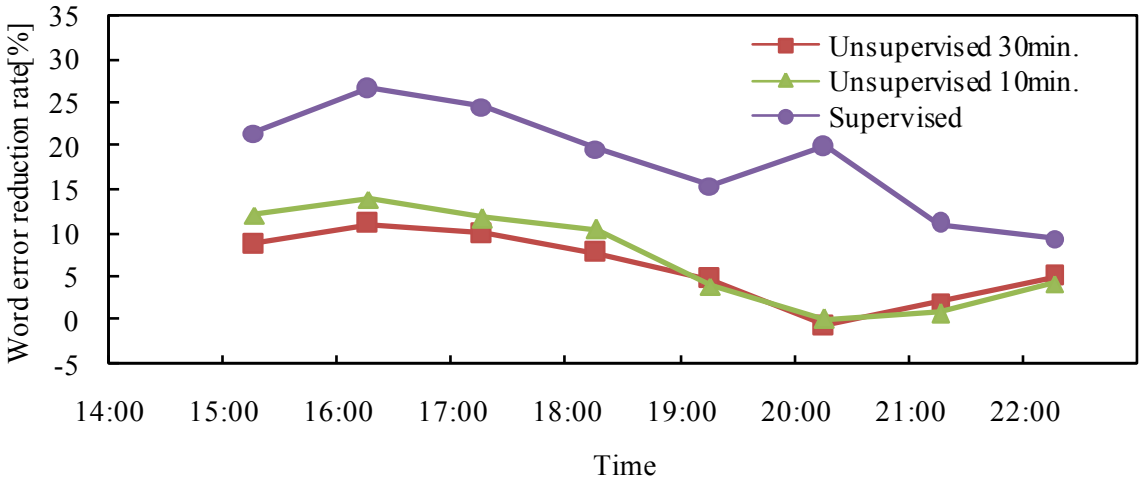

Fig. 6 Changes of word error reduction rates

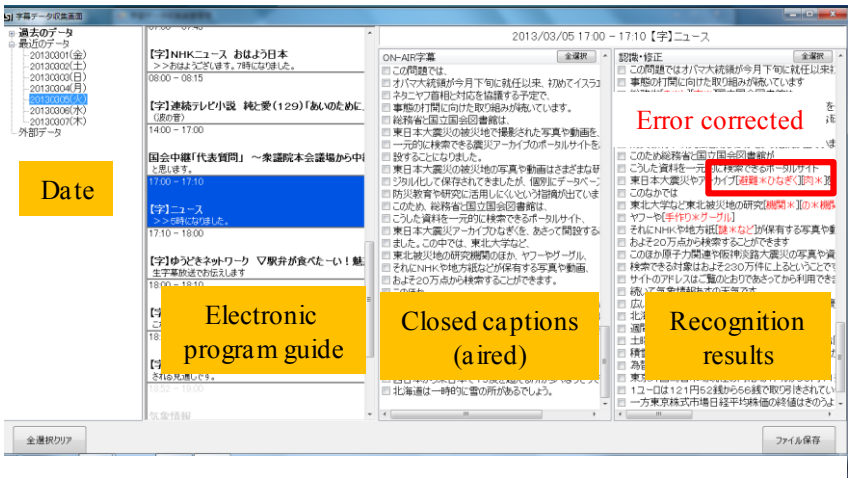

Fig. 7 Screen shot of data collection

Meanwhile, the broadcasting program around 19 o'clock reviewed the disaster with scenes edited from the news around 15 or 16 o'clock, so that the degradation around 19 o'clock was caused by the same background noise as the noise causing the degradation around 15 o'clock.

VI. Towards the Realization of Closed-captioning for Disaster Broadcasts

As mentioned above, the progressive LM adaptation method achieved a significant improvement for transcribing disaster broadcasts. To realize the method in our closed-captioning system, we developed a prototype device for real-time data collection. Fig. 7 shows a screen shot from the prototype. It archives closed captions and recognition results according to time slots of broadcasting programs, e.g. electronic program guides (EPGs). Users can select any sets of captions and recognition results as LM training texts. As the recognition results are displayed with error corrected results, where the color of the characters is changed to red for visibility, they can also gives priority for the supervised transcriptions by selecting colored characters. Meanwhile, automatic process from selecting data to adapting LM can be also conducted without manually selecting.

VII. Conclusion

For disaster broadcasting with closed captioning, we described an automatic LM adaptation method updating by using transcriptions as unsupervised training data, in addition to online manuscripts. We examined our proposed method for the disaster broadcasting occurred on March 11, 2011. Examination showed that the situation in terms of *lifeline* changed more rapidly than the other topics. In the evaluation of ASR performance, LM updating by using unsupervised training data reduced word error rate by 8.2% relatively compared with LM updating only by online manuscripts. During the beginning 4 hours, unsupervised adaptation in a short interval was desirable in performance. Meanwhile, during the next 4 hours, the online manuscripts became enough quantities to maintain ASR performance.

Our developed prototype for real-time data collection is capable of gathering not only recognition results but also manually error corrected results, which are supervised training data. In future work, we will intend to improve the ASR performance through the introduction of an efficient way of combining unsupervised and supervised training data.

References

[1] A. Kobayashi, K. Onoe, T. Imai, and A. Ando, "Time dependent language model for broadcast new transcription and its post-correction," *Proc. ICSLP*, pp. 2435-2438, 1998.

[2] H. Nanjo, and T. Kawahara, "Unsupervised language model adaptation for lecture speech recognition," *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.

[3] S. Novotney, R. Schwartz, and J. Ma, "Unsupervised acoustic and language model training with small amounts of labeled data," *Proc. ICASSP*, pp. 2210-2213, 2009.

[4] T. Imai, A. Kobayashi, S. Sato, H. Tanaka, and A. Ando, "Progressive 2-pass decoder for real-time broadcast news captioning," *Proc. ICASSP*, vol. 3, pp. 1559-1562, 2000.

[5] S. Homma, A. Kobayashi, T. Oku, S. Sato, and T. Imai, "Live closed-captioning system with robust speech recognition for a spontaneous spoken-style," *Proc. APSIPA*, pp. 450-453, 2000.

[6] T. Imai, S. Homma, A. Kobayashi, T. Oku, and S. Sato, "Speech Recognition with a Seamlessly Updated Language Model for Real-Time Closed-Captioning," *Proc. Interspeech-2010*, pp.262-265, 2010.

[7] T. Imai, S. Sato, A. Kobayashi, K. Onoe, and S. Homma, "Online speech detection and dual-gender speech recognition for captioning broadcast news," *IEICE Trans. Inf. & Syst.* E90-D, vol.8, pp. 1286-1291, 2007.