Automatic Classification of Usability of ASR Result for Real-time Captioning of Lectures

Yuya Akita*† Nobuhiro Kuwahara*[1] Tatsuya Kawahara*†
* School of Informatics, Kyoto University
† Academic Center for Computing and Media Studies, Kyoto University,
Sakyo-ku, Kyoto 606-8501, Japan

***Abstract*—As a support to hearing-impaired students in a classroom, real-time captioning and note taking using automatic speech recognition (ASR) have been investigated. However, even with ASR, editing by hand is needed to check and correct recognition errors and redundant spoken expressions in ASR results, and thus it often leads to delay in presenting captions. For efficient edit and quick presentation, we propose an automatic classification of ASR results in terms of usability as caption texts, and a presentation method based on the classification. In this study, we define the usability by syntactic correctness, errors and redundant spoken expressions in ASR results. Based on this definition, each unit of ASR results is classified into "valid," "invalid" or "to be checked," using hand-crafted rules and a machine learning framework. When presenting captions, "valid" input is presented promptly. "To be checked" input is manually edited, and then added to captions. We developed a real-time captioning system by incorporating the automatic classification method and the presentation method, and conducted a trial of this system in a university lecture.**

I. Introduction

In many countries, educational institutes are required to offer necessary support such as audio, visual, and/or physical aids to disabled students. As for hearing-impaired students, several efforts are conducted in classroom lectures to transform audio information into visual information. In Japan, typical manners of the transformation in classrooms are hand-written note taking and captioning with laptop computers, both by voluntary supporters. Since the number of notes or captions made by a single supporter is limited, two or more supporters are usually expected in a lecture, while there is difficulty in preparing a sufficient number of supporters. They should have some knowledge on the topics of the assigned lecture to make accurate transcripts. Particularly in university lectures, a variety of technical terms are often used, and these terms are difficult or impossible for non-experts to understand. It is always a severe problem to find supporters who have matched academic background, as well as fast writing or typing skill.

To alleviate this problem, automatic speech recognition (ASR) is promising. Some research groups, including our team, proposed automatic captioning of classroom lectures by using ASR [1], [2]. ASR has an advantage of fast and full transcription compared to a human. It can often be assumed that some materials of lectures such as textbooks, slides and audio recordings are available for topic adaptation of an ASR system. With an adapted ASR system, even technical terms can be recognized correctly. On the other hand, recognition errors and redundant spoken expressions such as repairs and hesitations are inevitable, and therefore these must be checked and edited by a human editor. As the edit is sequentially done over ASR results, delays often happen in presenting captions. In our previous work [1], the delay of captions was approximately nine seconds on average, and a half of the delay was caused in finding erroneous parts of ASR results.

For efficient edit and rapid presentation of ASR transcripts, we propose an automatic classification method of ASR results in terms of usability as captions. In this method, we consider whether an ASR result needs to be edited or not, and whether it should be presented as a caption or not. When a certain unit of ASR result is judged acceptable as it is, the unit is immediately presented. When a unit is judged useless, the unit is rejected. A human editor only edits ASR results which are judged to do so. By using this method, edit and presentation of captions are expected to be expedited.

In this paper, we first discuss the usability of ASR result, then describe how to perform classification in terms of the usability, together with experimental results over lecture data. We also illustrate methods of caption presentation using the proposed classification, which are followed by a report of trials in a real classroom.

II. Usability of ASR result

The most popular measure of usability of ASR result is the confidence measure score (CMS), for example, those calculated by using posterior probability of words in ASR hypotheses [3]. Several research works have also been conducted on automatic detection and correction of ASR errors, where discriminative frameworks are recently adopted [4], [5], [6], [7]. In spoken dialog systems (SDS), detection and rejection of meaningless inputs which should not be responded by a system are deployed. This framework is often implemented by using CMS of ASR result, while a machine learning framework was applied in [8]. The judgment in these frameworks was done basically in the viewpoint of correctness or reliability of ASR results.

In this work, we propose classification in terms of "usability as captions." This usability does not necessarily mean correctness or reliability of ASR result; we do not care about errors which are not harmful to understanding of captions by viewers, while we reject redundant spoken expressions even if these

[1]This work was done while he was a student at the School of Informatics, Kyoto University.

are recognized correctly. We also take account of necessity of edit by hand. That is, the usability is not represented in a simple binary form such as acceptance or rejection. Not only correctness of ASR result, but its grammatical importance and structure affect the necessity of edit. Thus we should not fully depend on CMS or labels on correctness of ASR result, although CMS is used as an important cue for classification.

III. Automatic classification method

In this work, by using morphological and acoustic features, we classify ASR result into one of three categories; *valid input*, *invalid input* and *to be checked*. We adopt a grammatical chunk as a unit for this classification. First, we automatically segment ASR result into chunks, then apply rule-based classification using syntactic information. As the rule does not consider ASR error and redundant spoken expressions, we further apply a framework of conditional random fields (CRF) to make a final decision.

A. Chunking of ASR result

Phrase-by-phrase editing and presentation is easier than the word-by-word basis, for both of editor and viewer. Thus we segment ASR input into grammatical chunks. Here, as we treat Japanese lectures, we employ a grammatical unit of Japanese language "*bunsetsu*" as a unit of classification, edit and presentation. This unit consists of one or compound content words accompanied by function words such as particle. Since lexical parsers are not designed for speech transcripts, we use a chunker based on support vector machines (SVM) to segment bunsetsu units. As an implementation of SVM-based chunker, we adopt "Yamcha" [9], and we train its model with "the Corpus of Spontaneous Japanese" (CSJ), which is a collection of academic lectures and public speeches, and hence it is expected to be effective and robust on speech transcripts.

B. Definition of labels

We classify each unit to "valid input," "invalid input" or "to be checked." For the classification, the following three factors are considered:

(a) Grammatical correctness
Grammatical structure of input unit is correct.
(b) Correctness of recognition of content words
All content words such as noun, verb, adjective and adverb are recognized correctly.
(c) Non-redundancy of expressions
The unit is not a redundant and unnecessary part as captions, for example, repetition or hesitation.

We define "valid input" is a unit which is syntactically correct and accurate regarding content words, and does not have redundant expressions in it. When a chunking result is not appropriate or there is any ASR error of a content word in it, this case is "to be checked." If any redundant expression is contained in a unit, this is defined as "invalid input." These can be described by combining (a), (b) and (c):

- *Valid input* which should be presented as captions: (a) and (b) and (c) are satisfied.
- *Invalid input* which should not be presented: not (c).
- *To be checked* which should be presented after manual edit and confirmation: the other cases.

C. Classification based on rules and CRF models

Regarding (a), we can define grammatical rules. For (b) and (c), we cannot make comprehensive description, and thus we adopt CRF to model these factors using training data. For each input, we test it with rules for (a) and two CRF models which correspond to (b) and (c) respectively, then combine the results of these tests to give a classification label, as described in the previous section.

In the rule-based classification, the decision is made based on the following pattern with part-of-speech (POS) tags of ASR result:

$$\mathrm{Dependent}^{*}\mathrm{Prefix}^{*}\mathrm{Independent}^{+}\mathrm{Suffix}^{*}\mathrm{Dependent}^{*} \quad (1)$$

where Dependent and Independent mean dependent word and independent word, respectively, '$*$' stands for zero or more repetition, and '$+$' stands for one or more repetition. When an input does not match the rule (1), it is immediately classified into "to be checked," as it has an improper structure as a bunsetsu unit. One exception is that a unit which consists of only dependent word(s) is classified into "invalid input" since it is useless as a caption.

We treat the classification on (b) and (c) as a labeling problem using three labels, and is modelled by CRF. The features considered in our CRF models are words in a unit, pronunciation of these words, CMS of ASR results, existence of pauses, POS tags and the number of content words in the unit. As for CMS, an average of CMS in a unit is used. We adopt CRF++[1] as an implementation of CRF. We need to prepare labeled data to train these CRF models. For (b), we make use of the CSJ. ASR is performed over all speech in the CSJ, and then the ASR results are aligned with corresponding manual transcripts to give reference labels (i.e., recognized word is correct or not) to each word in the ASR results. The number of words in the ASR results was 6.9M. Using the ASR results with several features and the reference labels, CRF model for (b) was trained. For (c), we prepared annotation with ASR results of lectures, as described later. This annotation was given by hand so that they met the definition of usability listed in the previous section.

This classification framework can be applied to a wide range of lectures, although models are trained with data of a specific domain. Factor (a) is a grammatical rule which is independent of vocabulary. For (b), the CRF model is trained with a large amount of lecture transcripts, thus the model is expected to cover various types of ASR errors in lectures. For (c), redundant spoken expressions observed in lectures depend on speakers, while the variety of these expressions are limited. Once we build models for (b) and (c), these are expected to be effective on lectures whose topics are different from those of training data.

[1] http://code.google.com/p/crfpp/

TABLE I
THE EVALUATION SET FOR AUTOMATIC CLASSIFICATION

Speaker	Acc. (%)	#Units	Proportion of labels (%)		
			Valid	Invalid	To be checked
Speaker A	76.5	1,622	59.1	7.8	33.1
Speaker B	77.9	1,537	67.2	5.3	27.5
Speaker C	81.6	1,508	72.3	5.2	22.5

D. Experimental evaluation

We conducted an experimental evaluation of the classification of usability with three academic lectures given at a symposium held by an institute of Kyoto University. The specification of these lectures is listed in Table I. Character accuracy of ASR results on this set (Acc. in Table I) is 77%–82%, and the number of bunsetsu units is 1,508–1,622. Among the units, 59%–72% are labeled as "valid input" and 23%–33% are labeled as "to be checked." Compared to these two labels, the number of "invalid input" is small. We performed automatic bunsetsu chunking over the lectures by the SVM-based method described in Section III-A, then annotated usability labels by hand.

First, we tested various combinations of features of CRF to clarify the effectiveness of each feature. The result is shown in Table II. Here, we performed 3-fold cross validation using single lecture as a test set and the others for training of CRF model. As an evaluation measure, we define classification accuracy as the number of correctly classified bunsetsu units divided by all units in the test set. As shown in Table II, result (4), in which word, pronunciation, POS tags and CMS were used as features, achieved the highest accuracy of 78.8%.

Then, we examine the classification result of each label. Table III shows the result by using the feature set of (4) in Table II. There are relatively a large number of misclassification of "valid input" to "to be checked," and "invalid input" to "valid input." However, these kinds of confusion are actually harmless in real captioning because check of "valid input" is easy, and contaminated "invalid input" does not affect the understanding of captions. In contrast, misclassification of "to be checked" to "valid input" should be reduced, as in this case an input, which should be checked and edited, is presented as it is. This misclassification was caused by ASR errors, because the factor (b), which considers ASR errors, has an impact in the classification of "to be checked." The recall rate of "invalid input" is low. This was caused because we had a small number of training samples for "invalid input."

IV. CAPTION PRESENTATION USING CLASSIFICATION RESULTS

In our previous real-time captioning system [1], ASR results are checked and edited one by one and then presented as captions. With the proposed classification method, "valid inputs" are presented without delay, and captions are revised by adding "to be checked" units after manual edit. We conducted a comparison of conventional presentation methods and the proposed method using automatic classification. The comparison was made on three methods, which were reviewed by hearing-impaired subjects.

TABLE II
CLASSIFICATION ACCURACY BY SETS OF FEATURES OF CRF

No.	Feature						Acc. (%)
	Word	Pron	POS	CMS	#Cont	Pause	
(1)	v	-	-	-	-	-	56.6
(2)	v	v	-	-	-	-	56.9
(3)	v	v	v	-	-	-	73.0
(4)	v	v	v	v	-	-	78.8
(5)	v	v	v	v	v	-	78.8
(6)	v	v	v	v	v	v	78.8

v: used, -: not used

Word: words in a unit, Pron: pronunciation of words, POS: POS tags of words, CMS: confidence measure score of words, #Cont: number of content words, Pause: existence of pause.

TABLE III
CLASSIFICATION RESULTS

Predicted / Reference	Valid input	Invalid input	To be checked	Sum	Ratio (%)	Recall (%)
Valid input	2,591	25	467	3,083	66.1	84.0
Invalid input	106	152	28	286	6.1	53.1
To be checked	337	9	952	1,298	27.8	73.3
Sum	3,034	186	1,447	4,667		

As features, words in a unit, pronunciation of words, POS tag and CMS were used.

A. Methods of caption presentation

Throughout the comparison, we basically use white characters on a black screen, unless otherwise described.

Method 1 is the conventional method used in our previous system [1] in which only edited ASR results are shown as captions. There is delay in presenting caption, while there are no errors in captions.

Method 2 is a combination of real-time ASR output and edited materials, where ASR results are first presented with gray characters, then these are overwritten by edited texts in white characters. This method is intended to improve response, without degrading correctness of captions. However, a viewer needs to look at two caption streams at the same time as the revision of captions is done asynchronously.

Method 3 is the proposed method, in which Method 2 is enhanced by incorporating the automatic classification. Each unit of ASR result is classified, then "valid input" is promptly presented, "invalid input" is discarded, and "to be checked" is forwarded to manual edit. When "to be checked" unit comes, each character is replaced with a hyphen '-' temporarily. The sequence is overwritten by edited texts afterwards in red characters. With this method, correct ASR results are immediately provided and erroneous units are shown after manual check and edit.

B. Experimental evaluation

For comparison of the methods, we prepared and used a simulated situation of lecture by using videos of real lectures, since it is difficult to prepare the same condition for all subjects when testing in real lectures. In this setting, we created ASR results and edited texts in advance, and synchronized them with the test video. The presenting timing of ASR results in Methods 2 and 3 is exactly same as the time of ASR output,

and that of captions is calculated based on an assumption that review of each unit takes 500ms and the editing operation takes 500ms per character.

We prepared excerpts of lectures listed in Table I. A ten-minute segment was extracted from each lecture, whose word accuracy is ranging from 78% to 82%, and thus we use three lectures. Each subject reviewed all lecture videos with different presenting methods. Here, the subjects were enrolled in, or recently graduated from a college or a university. They have experience of support by hand-written or PC-based note taking when attending classes. We showed the lecture video on a large screen in front of a subject, and captions were presented on a laptop PC.

The three methods were ranked by the subjects. As a result, Method 1 and Method 3 gained the highest rank on average. According to comments from the subjects, the advantage of Method 3 was quick presentation of captions, while Method 1 was the most familiar method for the subjects because usual note taking is provided to them in this manner. We considered the proposed Method 3 was supported by this experiment and decided to implement it in our real captioning system.

V. Trials in real lecture

A. System components

We developed a real-time captioning system for lectures using the proposed classification and presentation methods, combining with an ASR system. In this system, speech of a lecturer is directly input to the speech recognizer via wireless transmission. ASR result is then filtered to remove fillers, then automatically segmented into bunsetsu units. Each unit is classified by the proposed method as described in Section III-C. This classification is quick and its result is predicted immediately after each input. As a checking and editing tool, we adopt common software of typing, editing and presenting real-time captions for PC-based note taking in Japan. For a presentation device, we use a prompter which is a half-mirror screen and therefore a viewer can see both captions and scenes behind the prompter at a look.

B. Trials and results

Trials of the real-time captioning system were conducted in a course at the School of Informatics of Kyoto University. There was a hearing-impaired student enrolled in this class. For comparison, we conducted trials with two different captioning systems. In the first week, we tested a conventional Method 1 explained in Section IV-A, where all ASR results were checked by a human editor. In the second week, we tested the proposed system.

Both systems used the same ASR setting. As a decoder, we used Julius rev. 4.2.3, with the decoding parameters tuned to realize real-time output. Acoustic model was trained with the MPE criterion with academic lectures in the CSJ. VTLN and MLLR-based speaker adaptation were applied to the model. Language model was also trained with the CSJ transcripts, together with slide texts and transcripts of lectures given by the same lecturer in the past.

After the lectures, we had opportunities of interviewing the hearing-impaired student. The interviewing was done by a member of the support room for disabled students in Kyoto University, to avoid any bias from us. Regarding response of the systems, the proposed system looked faster than the conventional system. However, captions were not accurate, and thus they became a significant burden to the student. The reason was inaccurate ASR results; the character accuracy was 58.1% in the first lecture and 61.8% in the second lecture. Many misclassifications were caused and consequently much edit by the human editor was needed. When we did a rehearsal of the proposed system, in which character accuracy of ASR results was 69.8%, the editor reported that the editing work was smooth. The fact suggests that the operation of the proposed system requires character accuracy of 70% or above. The student also commented that caption presentation using a prompter is effective for better understanding of the situation in a classroom, as he could see both captions and the lecturer.

VI. Conclusions

We have proposed a classification method of ASR results in terms of usability as captions. The proposed method consists of a grammatical rule and a machine learning framework. We also designed a presentation method of captions using classification results, which was supported by hearing-impaired subjects in an experiment. Finally, we developed a real-time captioning system using the classification and presentation methods, and conducted trials in real lectures.

VII. Acknowledgments

The authors are grateful to Associate Prof. Yoko Yamakata of Kyoto University for kindly allowing us to conduct the trial of our captioning system in her lecture. This work was partly supported by JSPS Grant-in-Aid for Scientific Research #25730112.

References

[1] T.Kawahara, N.Katsumaru, Y.Akita, and S.Mori, "Classroom Note-taking System for Hearing Impaired Students using Automatic Speech Recognition Adapted to Lectures," In *Proc. Interspeech*, pp.626–629, 2010.

[2] P.Cerva, J.Silovsky, J.Zdansky, J.Nouza, and J.Malek, "Real-time Lecture Transcription using ASR for Czech Hearing Impaired or Deaf Students," In *Proc. Interspeech*, 2012.

[3] F.Wessel, R.Schluter, K.Macherey, and H.Ney, "Confidence Measures for Large Vocabulary Continuous Speech Recognition," *IEEE Trans. Speech & Audio Process.*, Vol.9, No.3, pp.288–298, 2001.

[4] Z.Zhou, H.Meng, and W.K.Lo, "A Multi-pass Error Detection and Correction Framework for Mandarin LVCSR," In *Proc. Interspeech*, pp.1646–1649, 2006.

[5] A.Allauzen, "Error Detection in Confusion Network," In *Proc. Interspeech*, pp.1749–1752, 2007.

[6] Z.Zhou, J.Gao, F.K.Soong, and H.Meng, "A Comparative Study of Discriminative Methods for Reranking LVCSR N-best Hypotheses in Domain Adaptation and Generalization," In *Proc. ICASSP*, Vol.1, pp.141–144, 2006.

[7] G.Kurata, N.Itoh, and M.Nishimura, "Training of Error-corrective Model for ASR without Using Audio Data," In *Proc. ICASSP*, pp.5576–5579, 2011.

[8] H.Majima, et al., "Spoken Inquiry Discrimination using Bag-of-words for Speech-oriented Guidance System," In *Proc. Interspeech*, 2012.

[9] T.Kudo and Y.Matsumoto: "Chunking with Support Vector Machines", In *Proc. NAACL*, 2001.