Post-error Correction in Automatic Speech Recognition Using Discourse Information

Sangwoo KANG, Ji-Hwan KIM*, Jungyun SEO
Department of Computer Science and Engineering, Sogang University, 121-742, Republic of Korea
**Corresponding author: kimjihwan@sogang.ac.kr*

***Abstract*—Overcoming speech recognition errors in the field of human–computer interaction is important in ensuring a consistent user experience. This paper proposes a semantic-oriented post-processing approach for the correction of errors in speech recognition. The novelty of the model proposed here is that it re-ranks the n-best hypothesis of speech recognition based on the user's intention, which is analyzed from previous discourse information, while conventional automatic speech recognition systems focus only on acoustic and language model scores for the current sentence. The proposed model successfully reduces the word error rate and semantic error rate by 3.65% and 8.61%, respectively.**

***Index Terms*—Post correction, Speech recognition, Re-ranking model, Analysis of user intention, Spoken language understanding, Spoken dialog system.**

I. INTRODUCTION

A spoken-language interface is convenient in many application environments, such as mobile information retrieval and car navigation. However, the inconsistent performance of Automatic Speech Recognition (ASR) systems makes it difficult to expand their application to advanced interactive systems such as service robots or ubiquitous computing. In a spoken-language interface, the key issue is in recovering from the reduced application-level performance, which is largely due to incomplete ASR outputs. Handling speech recognition errors is an essential part in the development of advanced interactive systems.

Interactive systems can handle sentences consisting of multiple words, and these inputs may include a user's implicit intention along with the discourse in interactions. Thus, in the performance criteria of the ASR systems employed in such systems, the level of correctness in expressing the user's intention from a recognized sentence is critical. However, most of such systems are considered one-best recognized sentence among ASR outputs. A key assumption is that the ASR output, which is based on a calculation of the user's intention, can be used to improve the performance of such system. For this reason, a semantic-oriented post-processing approach for the correction of speech recognition errors is proposed.

II. RELATED WORK

Some error-handling techniques were investigated for improving the quality of speech recognition as part of the post-error correction process. In previous studies of post-error correction, a rule-based approach was studied [1-2], where numerous lexical error patterns were collected and used to generate correction rules in order to minimize the number of errors. However, this approach has a disadvantage in that the corrections are limited to the collected lexical error patterns. Statistical approaches were also studied [3-5]. These methods applied a noisy channel model for error correction. The noise channel model has a distribution accounting for the possibility that an original word can be misinterpreted as an erroneous word. Such methods made use of the probability of lexical clues in error strings, such as incorrectly recognized words, the co-occurrence information extracted from the words and their neighboring words, and the tagged word bi-grams. Such statistical approaches depend on the quantity and the quality of these collected error patterns. However, collecting a sufficient amount of error patterns is an intensive and time-consuming affair.

The most recent studies focus on improving the performance of an ASR system in terms of the Word Error Rate (WER) [6-8]. However, this is not considered the end result in application systems using a spoken-language interface. Presented herein is a new method that improves the quality of speech recognition in application systems. In addition, this method does not require the collection of error patterns. The basic idea of the proposed model is that the understanding of utterances is influenced by discourse information present in human-to-human dialog. In this light, the proposed model re-ranks the n-best hypotheses of ASR output by using this discourse information. To measure the performance in the proposed system, the Semantic Error Rate (SER) is used in addition to the WER.

Each n-best hypothesis is analyzed in a semantic form by a Spoken Language Understanding (SLU) module [9]. A user's intention is defined by the semantic form, which includes predicted meanings consisting of a Speech Act (SA), a Concept Sequence (CS), and a Named Entity (NE) [10-13]. SA represents the general intention expressed in an utterance, while CS captures the semantic focus of an utterance. NE is defined as any domain-specific proper noun. When discourse information composed in semantic form is given, the proposed model re-ranks the n-best hypotheses by generating their respective probabilities.

In order to assign the appropriate weights to features in the re-ranking model, a feature-weighting scheme based on Support Vector Machines (SVMs) [14-15] is used. This scheme clearly and automatically assigns optimal weights. The proposed model performs effectively in an interactive

This work was supported by the IT R&D program of MOTIE/MSIP/KEIT.[10041678,The Original Technology Development of Interactive Intelligent Personal Assistant Software for the Information Service on multiple domains] and this research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(No. NRF-2013R1A1A2010190)

Digital Object Identifier 10.4316/AECE.2014.02009

TABLE I. AN EXAMPLE OF UTTERANCES ALONG WITH THEIR CORRESPONDING SPEECH ACTS, CONCEPT SEQUENCES AND NAMED ENTITIES: (S: A SYSTEM, U: A USER, ITALIC MEANS NAMED ENTITY)

Utterance	Speech Act	Concept Sequence
U: Hello.	Greeting	NULL
S : May I help you?	Opening	NULL
U: Tell me the *tomorrow* schedule.	Request	Timetable-search
S : You have an appointment with *Kildong Hong* at *eleven a.m.*	Response	Timetable-search
U: We changed the appointment.	Inform	Timetable-modify
S : What is changed?	Ask-ref	Timetable-modify
U: The appointment date was changed.	Response	Timetable-modify-date
S : When is the changed date?	Ask-ref	Timetable-modify-date
U: It's *December five*.	Response	Timetable-modify-date

system by using spoken language because it is similar to the process of understanding that occurs in human-to-human dialog. Furthermore, when the ASR system is extended to new fields, additional costs are not incurred in collecting new error patterns.

This paper is organized as follows. In the next section, an overview of the proposed system is given. Subsection III.A explains both the SLU model and the re-ranking model for ASR post-correction, while subsection III.B explains the feature-weight scheme. Section IV details an experiment evaluating the different models, and in the final section, conclusions are presented.

III. ASR Post-Correction Model Using Discourse Information

The proposed model consists of two steps, as shown in Fig. 1. In the first step, the SLU model analyses n-best hypotheses generated by ASR outputs to construct semantic forms. Each semantic form is the user's intentions as interpreted from each hypothesis. In the second step, the re-ranking model orders the hypotheses using the generation probabilities of each semantic form when discourse information is included; discourse information is also composed of a semantic form.

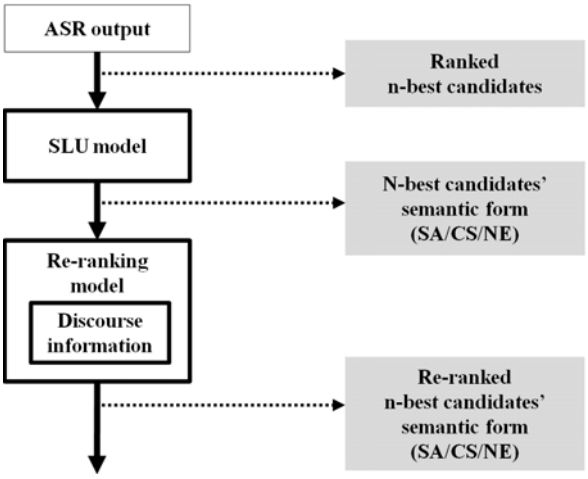

Figure 1. Overall system procedure

A. SLU model for analysis of n-best hypotheses

The goal of SLU is to construct a semantic form from a hypothesis and a user's intention is defined by a semantic form. The elements of semantic form consist of SA, CS, and NE. SA is generally domain-independent and is used to represent the main intention of the user in an utterance. A CS is created compositionally from an inventory of domain-dependent concepts. It may contain zero or more concepts. SA and CS form a domain action that is used to represent a user's intention to achieve some domain-dependent activity. In addition, proper nouns, such names of a location or a person, provide important clues for deriving the meaning of an utterance. To recognize domain-specific proper nouns, five NE types are defined: "date," "time," "place," "person," and "content." Table I shows examples of elements of a semantic form.

To determine the elements of the semantic forms in each hypothesis, the SLU model [16] was developed. The SLU model for determining SA can be formally defined as (1). Let $SA(h_i)$ and $CS(h_i)$ denote SA and CS of the i^{th} hypothesis (h_i) in the current utterance. The sentential feature set is represented by F_i: lexical features (content words annotated with Parts-of-Speech (POSs)) and POS features (POS bi-grams of all words in an utterance). The contextual feature $p.SA$ is determined by the SA of a previous utterance that is always a system utterance. The values of $P(SA|F_i)$ and $P(SA|p.SA)$ can then be approximated using these features. Similarly, CS can be determined as shown in (2).

$$SA(h_i) \approx \underset{SA}{argmax}\, P(SA \mid F_i) P(SA \mid p.SA) \quad (1)$$

$$CS(h_i) \approx \underset{CS}{argmax}\, P(CS \mid F_i) P(CS \mid p.CS) \quad (2)$$

The NE recognition model used in [17] was applied; this model uses a modified Hidden Markov Model (HMM) based on character n-grams.

The re-ranking model orders the semantic forms of each hypothesis by their probabilities, calculated from (3), where $SF(h_i)$ denote the semantic form of the h_i in the current utterance. $SA(h_i)$, $CS(h_i)$, and $NE(h_i)$ are elements of $SF(h_i)$. This equation gives the generation probability of each $SF(h_i)$ when the semantic form of a previous utterance ($p.SF$) is given. Consequently, $P(SF(h_i)|p.SF)$ in (3) can be replaced with $P(SA(h_i),CS(h_i)|p.SA,p.CS)$, where the elements of $SF(h_i)$ are determined by (2) and (3) and the NE recognition model. $P(SA(h_i)|p.SA)\times P(CS(h_i)|p.CS)$ is obtained by assuming the SA and CS to be independent. Finally, assuming that the $p.SA$ and $p.CS$ affect the type of NE (NE^{type}) in the current utterance, each $SF(h_i)$ is re-ranked according to the probability determined using (3).

TABLE II. WER AND SER OF DIFFERENT MODELS (%)

Models	Probability models for re-ranking	SER	WER
Baseline model	Top-1 hypothesis ASR of output	16.83	34.83
Model 1	$P(SA(h_i)\|p.SA) \times P(CS(h_i)\|p.CS)$	17.12	34.00
Model 2	$P(NE^{type}(h_i)\|p.SA, p.CS)$	16.24	34.33
Model 3 (proposed)	$P(SA(h_i)\|p.SA) \times P(CS(h_i)\|p.CS) \times P(NE^{type}(h_i)\|p.SA, p.CS)$	15.88	32.33
Model 4 (proposed)	*Applying a feature weighting scheme to Model 3*	15.78	31.83

$$P(SF(h_i)\mid p.SF) := P(SA(h_i)\mid p.SA) \times P(CS(h_i))\mid p.CS) \times P(NE^{type}(h_i)\mid p.SA, p.CS) \quad (3)$$

B. Feature-weighting scheme using SVM

A feature-weighting scheme is used to approximate the optimal degree of influence of individual features. The general approach for estimating feature weights is to use empirical methods; however, this approach is *ad hoc* and does not guarantee a credible and optimal value. Equation (4) represents a form that includes the degrees of influence of three terms from (3), where w_1, w_2, and w_3 represent the respective weights of the features.

$$\log P(SF(h_i)\mid p.SF) := w_1 \log P(SA(h_i)\mid p.SA) + w_2 \log P(CS(h_i)\mid p.CS) + w_3 \log P(NE^{type}(h_i)\mid p.SA, p.CS) \quad (4)$$

A discriminant model [18] is introduced to estimate the degrees of influence of individual features by using a training set. This method can efficiently decide the optimal degrees of influence of terms in (4) using SVM learning. For SVM learning, the feature that consists of the three terms in (3) is used, where the values of each term are calculated from a corpus. This method also requires a correct and incorrect training set for h_i, because SVM is designed for binary classification. A Text-to-Speech (TTS) system is used to collect the voice data for this system. The research version of VOICEWARE's VoiceText™ was employed for the experiments. Voice data is automatically collected by using VoiceText™, and this data is then used as input for the ASR system. The ASR outputs, which are divided into correct and incorrect data, are then used as inputs for the SLU system. The SVM automatically estimates the degrees of influence of individual terms. Thus, the higher the value of the SVM, the greater is the accuracy. Finally, the proposed model re-ranks the n-best hypotheses of ASR output by the value of the SVM.

IV. Experiment and Results

To evaluate the performance of the proposed model, a tagged corpus was used. This corpus contains 6,953 pairs (6,353 pairs for training and 600 for testing) of system/user utterances in the schedule management domain. This corpus is annotated with a tag set consisting of 12 SA, 43 CS, and 5 NE tags. The Korean continuous speech recognizer described in [19] was used. The experimental platform included a personal computer with an 8 GB RAM, I5(2.7GHz) CPU and C++ language. We implemented the SA and CS analyzer to adopt the basic idea of Kim's model [12]. Each analyzer used SVM^{light} with a linear kernel; SVM^{light} is an SVM implementation of Vapnik [14], and consists of multiple SVM classifiers for multi-class classification. Further, a dictionary-based NE recognizer was employed.

The WER, SER, and Error Reduction Rate (ERR) were employed as criteria for evaluating the performance of the proposed system; the semantic error means that the top-1 hypothesis' semantic form does not match exactly up with a gold standard for transcripts. For experimental purposes, these three metrics were defined as shown in (5–7):

$$WER = \frac{the\ \#\ of\ word\ [substitutions + deletions + insertions]\ errors}{the\ \#\ of\ words} \quad (5)$$

$$SER = \frac{the\ \#\ of\ semantic\ structure\ [SA + CS + NE]\ errors}{the\ \#\ of\ utterances} \quad (6)$$

$$ERR = \frac{the\ error\ rate\ of\ the\ b.model - the\ error\ rate\ of\ the\ p.model}{the\ error\ rate\ of\ the\ b.model} \quad (7)$$

$b.model = baseline\ model\ ,\ p.model = proposed\ model$

TABLE III. ERRS OF DIFFERENT MODELS

Models	ERR of WER (%)	ERR of SER (%)
Model 1	-1.72	2.38
Model 2	0.84	1.44
Model 3 (proposed)	3.09	7.18
Model 4 (proposed)	3.65	8.61

Table II and III show the WER/SER of different models and the ERRs of WER/SER. In the performance of a baseline model, the top-1 hypothesis ASR of the output is compared to the reference transcription. Three types of re-ranking models were defined for various evaluations. Models 1 and 2 use $P(SA(h_i)|p.SA) \times P(CS(h_i)|p.CS)$ and $P(NE^{type}(h_i)|p.SA,p.CS)$, respectively, as probability models for re-ranking. Model 3 applies $P(SA(h_i)|p.SA) \times P(CS(h_i)|p.CS) \times P(NE^{type}(h_i)|p.SA,p.CS)$ using (3): the combination of Models 1 and 2. Finally, Model 4, as the proposed model employs a feature-weighting scheme to $P(SA(h_i)|p.SA) \times P(CS(h_i)|p.CS) \times P(NE^{type}(h_i)|p.SA,p.CS)$. Models 1, 2, and 3 have an identical feature weight. However, Model 4 applies the feature weighting scheme that have the optimal degree considering the influence of each feature. Each feature weight is decided efficiently by the machine learning method (details in Section *III.B*).

Model 1 shows a positive result for the ERR of SER because Model 1 uses the semantic-oriented features *p.SA* and *p.CS*; however, Model 1 shows a negative result for the ERR of WER. This shows that the absence of NE^{type} as a content word has a negative effect on the WER. Model 2 has

a positive result for both ERRs of WER and SER; further, the ERR of SER of Model 2 is smaller than that of Model 1. This means that *p.SA* and *p.CS* are more important semantic-oriented features than NE^{type} in SER. The proposed Model 3 and 4 use all semantic features, *p.SA*, *p.CS*, and NE^{type}. It has the best results compared to all the other models. In particular, the SER for Model 3 is reduced by 7.18%. In addition, a feature-weighting scheme using SVM reduces ERRs further because the ERRs of Model 4 are slightly higher than Model 3.

In all models, it is confirmed that the proposed method is highly effective in reducing SER, and it is further shown that the proposed system also contributes to reducing WER. In addition, a significant increase in the performance of an interaction system using a spoken-language interface is also expected.

V. Conclusions

In this paper, a new post-error collection model is proposed for handling erroneously recognized outputs. This model re-ranks the n-best hypotheses of ASR using discourse information to improve the performance of an interactive system. It was found that this re-ranking model effectively reduces the error rates. In practice, this model reduces SER to approximately 8.61%, which implies that the model is effective in finding a hypothesis closest to the user's intention. Moreover, when the interactive system is extended to new fields, the collection of error patterns would require further efforts and costs to customize the system to those fields.

The future research aims to find efficient way to implement intelligent personal assistant in ubiquitous computing [20], which presents significant technical challenges. It will be analyzed in terms of the following four characteristics of which implementation has: 1) It should be lightweight in terms of computation and memory requirements, because it runs on simple devices with limited resources. 2) It requires robustness to corruption in environment. There is clear distinction in acoustic features between the acoustic model training data and the dialogues. 3) It requires easy extension to include personal information. The named entities of users' interest are different from the context. This personalized information would greatly enhance the performance of speech recognition from the user's point of view. 4) It estimates suitable probabilities for unseen sequences and copes well with this data scarcity problem. As speech recognition in ubiquitous computing devices involves a lot of personal data, it costs a great deal to gather a sufficient corpus. Thus, it is very difficult to collect such amount of corpus as it is required to measure the exact frequency of a word or syllable sequences.

References

[1] S. Kaki, E. Sumita, H. Iida, "A Method for Correcting Errors in Speech Recognition Using the Statistical Features of Character Co-occurrence," in Proc. of Association for Computational Linguistics, pp. 653-657, 1998. [Online]. Available: http://dx.doi.org/10.3115/980845.980954

[2] R. Lopez-Cozar, Z. Callejas, "ASR Post-Correction for Spoken Dialogue Systems based on Semantic, Syntactic, Lexical and Contextual Information," Speech Communication, vol. 50, no. 8-9, pp. 745-766, 2008. [Online]. Available: http://dx.doi.org/10.1016/j.specom.2008.03.008

[3] J. Allen, B. W. Miller, E. K. Ringger, T. Sikorski, "A Robust System for Natural Spoken Dialog," in Proc. of Association for Computational Linguistics, pp. 62-70, 1996. [Online]. Available: http://dx.doi.org/10.3115/981863.981872

[4] E. Ringger, J. Allen, "A Fertility Channel Model for Post Correction of Continuous Speech Recognition," in Proc. of International Conference on Spoken Language Processing, pp. 897-900, 1996. [Online]. Available: http://dx.doi.org/10.1109/ICSLP.1996.607746

[5] M. Jeong, G. G. Lee, "Improving Speech Recognition and Understanding using Error-Corrective Reranking," ACM Transactions on Asian Language Information Processing, vol. 7, pp. 2:1-2:26, 2008. [Online]. Available: http://dx.doi.org/10.1145/1330291.1330293

[6] T. Hazen, T. Burianek, J. Polifroni, S. Seneff, "Recognition confidence scoring for use in speech understanding systems," Computer Speech and Language, vol. 16, no. 1, pp. 49-67, 2002. [Online]. Available: http://dx.doi.org/10.1006/csla.2001.0183

[7] T. Baumann, M. Atterer, D. Schlangen, "Assessing and improving the performance of speech recognition for incremental systems," in Proc. Of Association for Computational Linguistics, pp. 380-388, 2009. [Online]. http://dx.doi.org/10.3115/1620754.1620810

[8] C. Clavel, G. Adda, Cailliau, M. Garnier-Rizet, A. Cavet, G. Chapuis, S. Courcinous, C. Danesi, A. Daquo, M. Deldossi, S. Guillemin-Lanne, M. Seizou, P. Suignard, "Spontaneous speech and opinion detection: mining call-centre transcripts," Language Resources and Evaluation, 2013. [Online]. Available: http://dx.doi.org/10.1007/s10579-013-9224-5.

[9] J. Vilaneau, J. Y. Antoine, "Deeper Spoken Language Understanding for Man-machine Dialogue on Broader Application Domains: A Logical Alternative to Concept Spotting," in Proc. of Workshop on Semantic Representation of Spoken Language, pp. 50-57, 2009. [Online]. Available: http://dx.doi.org/10.3115/1626296.1626303

[10] H. Lee, H. Kim, J. Seo, "Efficient Domain Action Classification using Neural Networks," Lecture Note in Computer Science, vol. 4233, pp. 150-158, 2006. [Online]. Available: http://dx.doi.org/10.1007/11893257_17

[11] H. Kim, "A Dialogue-based NLIDB System in a Schedule Management Domain: About the Method to Find User's Intentions," in Proc. of conference on Current Trends in Theory and Practice of Computer Science, pp. 869-877, 2007. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-69507-3_75

[12] D. Kim, H. Lee, C. Seon, H. Kim, and J. Seo, "Speakers' Intention Prediction Using Statistics of Multi-level Features in a Schedule Management Domain," in Proc. of Association for Computational Linguistics on Human Language Technologies, pp. 229-232, 2008. [Online]. Available: http://dx.doi.org/10.3115/1557690.1557756

[13] H. Kim, C. Seon, J. Seo, "Review of Korean speech act classification: machine learning methods," Journal of Computing Science and Engineering, vol. 5, no 4, pp. 288–293, 2011. [Online]. Available: http://dx.doi.org/10.5626/JCSE.2011.5.4.288

[14] V. Vapnik, The Nature of Statistical Learning Theory. Springer Verlag, 1995. [Online]. Available: http://dx.doi.org/10.1007/978-1-4757-2440-0

[15] L. Bottou, C. Cortes, J. Denker, H. Drucker, I. Guyon, L. Jackel, Y. LeCun, U. Muller, E. Sackinger, P. Simard, and V. Vapnik, "Comparison of Classifier Methods: A Case Study in Handwritten Digit Recognition", in Proc. of International Conference on Pattern Recognition, vol. 2, pp. 77-82, 1994. [Online]. Available: http://dx.doi.org/10.1109/ICPR.1994.576879

[16] S. Kang, H. Kim, J. Seo, "A Reliable Multidomain Model for Speech Act Classification," Pattern Recognition Letters, vol. 31, no 1, pp. 71-74, 2010. [Online]. Available: http://dx.doi.org/10.1016/j.patrec.2009.08.013

[17] C. Seon, H. Kim, J. Seo, "Efficient Appointment Information Extraction from Messages in Mobile Devices with Limited Hardware Resources," Pattern Recognition Letters, vol. 32, no 2, pp. 127-133, 2011. [Online]. Available: http://dx.doi.org/10.1016/j.patrec.2010.09.029

[18] R. Nallapati, "Discriminative Models for Information Retrieval," in Proc. of SIGIR, pp. 64-71, 2004. [Online]. Available: http://dx.doi.org/10.1145/1008992.1009006

[19] K. Lee, M. Chung, "Morpheme-Based Modeling of Pronunciation Variation for Large Vocabulary Continuous Speech Recognition in Korean," IEICE Transaction on Information and Systems, vol. E90-D, no. 7, pp. 1063-1072, 2004. http://dx.doi.org/10.1093/ietisy/e90-d.7.1063

[20] M. Lee, D. Han, "Ubiscript: A Script Language for Ubiquitous Environment," Journal of Computing Science and Engineering, vol. 5, no 2, pp. 141–149, 2011 [Online]. Available: http://dx.doi.org/10.5626/JCSE.2011.5.2.141