Clarification Question Generation for Speech Recognition Error Recovery Using Monolingual SMT

Dong Yu
International R&D Center for Chinese Education, Beijing Language and Culture University,
Beijing, 10083, China
E-mail: yudong@blcu.edu.cn

***Abstract* - Clarification dialogue is an efficient and direct way of handling speech recognition errors in speech interface applications. In this paper we present a new approach to Clarification Question (CQ) generation. Monolingual phrase-based SMT (PB-SMT) framework is introduced to generate robust and flexible CQs. A parallel corpus from simulated error to manually annotated CQ is established and used for training the model. A new type of generalized phrase pair is expanded from conventional translation phrase table. Combining both generalized and conventional phrase pairs, a two-step decoding process is carried out to generate CQs. Both manually and automatic metrics are used to evaluate the quality of generated CQs. Experimental results show that our method can effectively generate reasonable CQs form miss-recognized utterances, and generated CQs can be used to prompt a clarification dialogue for error handling.**

Index Terms – clarification dialogue, question generation, speech recognition error recovery, monolingual SMT

I. INTRODUCTION

In open domain speech interface application, a robust error handling method for speech recognition process is quite essential. People in human-human conversation try to clarify the major cause of mishearing and misunderstanding to correct them. As an imitation of this phenomenon, Clarification Dialogue (CD) can be applied to speech recognition process for error handling. As shown in Fig. 1, user utterance is recognized and sent to the CD module, miss-recognized utterance are clarified through human-machine interaction and the result is used as final recognition output. There are two key issues for prompting such clarification dialogue: (1) decide if the utterance needs to be clarified; (2) generate an appropriate CQ to prompt the dialogue. In this paper, we focus on the second issue.

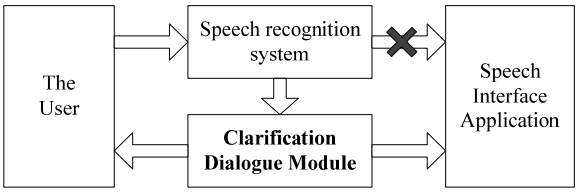


Figure 1. Clarification Dialogue in speech interface application.

Using clarification dialogue for error recovery in human-machine interaction has been studied by many researchers. Current researches can be divided into two categories. The first one is spoken dialogue system (SDS) based clarification, including works of Schlangen [1], Purver [2], Skantze [3]. These works are task oriented, where CD is aimed at restricting human-machine interaction to the task domain, and out-of-domain input will be treated as misunderstanding. The second one is specific task based clarification, including works of Lewis et.al [4], Misu et al. [5] and Sangkeun et al. [6]. These works require that the task knowledge has a well defined structure, or it can be represented into a certain form. Both kinds of CD metrics have limitations in dealing with speech recognition errors in general domain applications. On one hand, there is no definite task knowledge in general domain application, so that it is impossible to build a NLU model or a task knowledge structure. On the other hand, both kind of metrics use previously designed expressions for prompting clarification dialogue, so lack of flexibility and cannot adapt to general domain.

To overcome these drawbacks, this paper proposes a new approach to automatic CQ generation for handling speech recognition errors. The problem is seen as a monolingual SMT task [7]. A simulated miss-recognition database is established. PB-SMT [8] framework is introduced as CQ generator. Mainly two improvements are proposed to assure the PB-SMT framework: (1) a new type of generalized phrase pair is introduced for CQ generation model training; (2) a two-step decoding method is proposed for CQ generation. The whole process is for open domain and is task independent. CQ generated from PB-SMT framework is more natural and flexible than limited clarification strategies. Both automatic and manual evaluation metrics are used to evaluate the quality of generated CQs.

This paper is organized as follows: Section II studies features of miss-recognized utterance and describes the data annotation task. Section III proposes our method of CQ generation. Section IV gives experimental results and analysis. Finally our conclusions are drawn in Section V.

II. CLARIFICATION DATA ANNOTATION

A. Speech Recognition Error Analysis

In this work, totally 2,600 spoken language utterances in Chinese are recognized following [9] The system achieves an average 85.3% WER in one-best outputs. Totally 265 utterances are miss-recognized. We classify all miss-recognized utterances into three types according to their error severity. Table I shows the criterion and the proportion of each type in our database.

Both trivial and serious errors have fixed clarification strategies. For trivial error, explicit confirmation is used to confirm the error [1]; for serious error, the system directly asks the user to rephrase his/her utterance. However, as shown in Table I, partial error take account for more than 3/4 of all errors. Rephrase strategy is not suitable for partial error because part of the utterance is correct recognized, and confirmation strategy is also not suitable as that the incorrect part cannot be ignored. In

Published by Atlantis Press, Paris, France.

this situation, a CQ-lead clarification dialogue can be the most appropriate way to handle it. The user can answer the question to clarify the incorrect part of the utterance. In this work, we focus on how to generate such CQ for partial errors in a robust and flexible way.

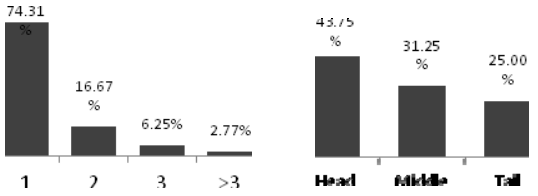
TABLE I. CRITERION AND PERCENTAGE OF EACH ERROR CATEGORIES

Category	Criterion of the Category	Proportion
Trivial Error	Only one little error exists, and it could be ignored.	15.47%
Partial Error	Only one error exists, some of information is lost.	78.52%
Serious Error	One or more error exists, has little useful information.	6.01%

B. CQ Annotation

Although real speech recognition process can provide a small amount of miss-recognition samples, the size is far from sufficient for the CA to learn human's clarification expressions. So this work simulates recognition errors from text corpus to establish a simulated error database. In this way, a large size of simulated error samples can be obtained. Furthermore, through manual annotation, a corresponding clarification request data set can be established for training machine learning techniques. Stuttle et.al [11] proposes an approach to ASR data simulating using WFST. However such model is not suitable in general domain. In this work, we simulate recognition errors by considering following two features: (1) Number of errors: a consecutive error word sequence is treated as one error; (2) Error position: the error may be occurred in Head, Middle, or Tail of the utterance.

In order to make the simulated data more realistic, we count distributions of two features in real recognized data, and apply them in error simulation process. As shown in Fig. 2. We use IWSLT 2005 released Chinese training data for error simulation. For each sentence in the corpus, a number of consecutive words are selected according to above two distributions. Selected words are hidden and replaced by an "Err" symbol to simulate a mishearing fragment.



(a) Number of error words (b) Error position

Figure 2. Distributions of two features. (a)is computed according to error word number of each utterance; in(b), three different of error types are concerned, "Head" means the error is occurred at the beginning of the utterance.

Human annotators are asked to annotate the simulated utterances. A CQ is generated for each partial error utterance. Trivial errors and serious errors are not required to tag CQs. Ambiguity and misunderstanding of the utterance is ignored during annotation. The generated CQs must include an interrogative word related to the error position, and at least one correct recognized word. Table II shows statistics of the annotation data. As shown in the table, the proportion of partial error in simulated data is close to that of real miss-recognized data.

TABLE II. STATISTICS OF ANNOTATION DATA

Statistics	Data
Number of utterances in corpus	20,000
Number of partial error utterances	15,772
Percentage of partial error	77.86%
Average length of partial error utterances	7.13
Average length of CQs	4.78
Vocabulary size	6,213

III. THE CQ GENERATION MODEL

A. Using Monolingual PB-SMT for CQ Generation

PB-SMT is a well known SMT method. It uses aligned sequences of words, named phrase pair, to build translation model. Such a model normally finds a best translation $\hat{e}$ of a string in source language *f* to a string in target language *e* by combining a translation model $p(f|e)$ with a language model $p(e)$:

(1)

New target language translations can be constructed by connecting phrase pairs in different ways. PB-SMT framework provides a general way of mapping between two kinds of texts, not only two languages. So if we treat miss-recognized utterances as "source language", and clarification questions as "target language", a PB-SMT model can be established between the two.

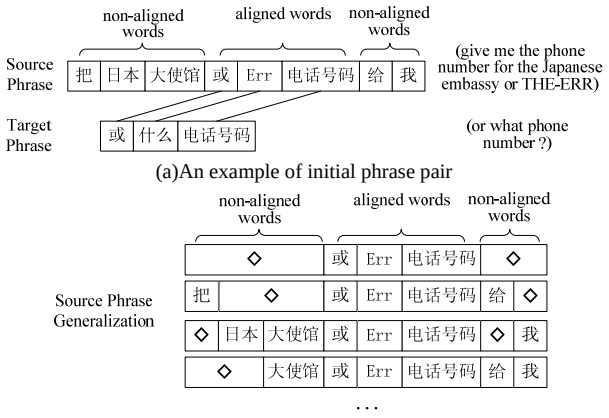
There are many advantages of using PB-SMT framework for CQ generation. On one hand, PB-SMT is a task independent method; its outputs are only related to the training data. On the other hand, PB-SMT has the capability of generating flexible and natural translation results. Moreover, the generated CQ is closely related to speaker's utterance, so it is very easy for the speaker to understand and to reply it. However, CQ generation is different from translation. As shown in Table.2, the average length of CQ is much less than that of miss-recognized utterance. It means that human only focus on the error point, and the irrelevant information is ignored when generate a CQ. Moreover, one CQ can be used to clarify a lot of miss-recognized utterances, if these utterances have the same error context. Therefore the generation model needs to be generalized, and error irrelevant information has to be ignored by the model. In this work, improvements on phrase table extraction are carried out to do this, as described in the next section.

B. Phrase Pair Extraction for CQ Generation

Extraction of phrase pair from bilingual corpus is the heart of PB-SMT framework. As in most conventional PB-SMT systems, GIZA++ is used to perform the word alignments which are then used to generate phrase alignments. Considering the CQ generation is a monolingual translation, obviously omitted alignments are added in this step. Phrase table is then established according to it. On the other hand, word alignments are used to distinguish error relevant words and error irrelevant words. This is done by a separate phrase generalization process.

The phrase extraction and generalization process can be divided into following steps:

- Extracting initial phrase table from clarification data follows the method of [8].
- Source phrase with an "Err" symbol is used for generalization. All aligned words are uses as skeleton of the phrase. Non-aligned words are selectively substituted by a placeholder "◇" for generalization. Consecutive non-aligned words are seen as one cell and substituted by a placeholder. At most two placeholders are allowed in one source phrase. All possible generalization situations must be traversed.
- All newly constructed source phrases are aligned to the original aligned target phrase.



(b) Examples of phrase generalization.

Figure 3. An example of phrase pair extraction

In this way, a number of new source phrases are constructed. Fig. 3 shows an example of this process. The advantage of generalized phrase is obviously. First of all, placeholder in source phrase provides a generalization capability. The generalized phrase can be applied to more error situations. Meanwhile, we need not to change the aligned target phrase as that only non-aligned words are used for substitution. Secondly, the phrase used for generalization must have an Err symbol. In this case, non-aligned words are also error irrelevant words, and are ignored by the aligned target phrase. So it is reasonable to believe that if these non-aligned words are changed, the target phrase is still effective when generating a CQ. Therefore, by using generalized phrases, we only need to focus on the skeleton of miss-recognized utterance; error irrelevant words can be automatically ignored.

Following conventional PB-SMT models, we use phrase translation weights and lexical weights to measure phrase translation quality. Considering that there may be unseen phrase in generalized source phrases, an "add one" smooth strategy is applied.

C. The Two-step Decoder for CQ Generation

As in most of PB-SMT systems, in this work beam-search decoder is used to find the optimal outputs. Different from conventional translation process, generalized phrase has higher priority than un-generalized phrase in CQ generation task. It means if a generalized phrase can be implemented on an input utterance, un-generalized phrase can be ignored, no matter what a high translation score it has. So the decoding process of CQ generation is divided into two steps. In the first step, only generalized phrases are used for decoding. The Decoder-I selects all generalized phrases that can be implemented on the input utterance as initial translation candidates. Each placeholder is filled according to the input. The best initial translation candidates that can cover all input utterance is the optimal CQ. If there is no initial translation candidate can cover all input utterance, the Decoder-II is started up to translate all uncovered words. Un-generalized phrases combined with all generalized phrases selected by Decoder-I are used in this step, its output is the optimal CQ. Fig.4 shows the framework of the two-step decoding process.

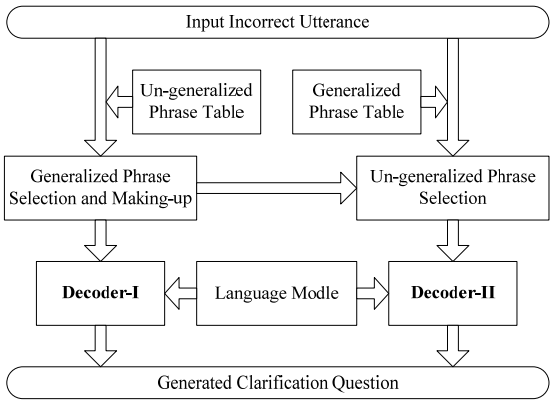


Figure 4. An overview of the two step decoding process

The Decoder-I tries to use one generalized phrase to clarify the error. It can generate a natural and compact CQ because the selected phrase only focus on the error point and all irrelevant words are ignored. The Decoder-II can be seen as a back off of Decoder-I. Uncovered words are translated because they may be error relevant and cannot be ignored. The two-step decoding process takes full advantage of two types of phrase pairs. It generates perfect CQs if the error is definitely. In more complicated cases, it can retain more information and generate reasonable CQs.

IV. EXPERIMENTAL RESULTS

All experiments are carried out in Chinese spoken language domain. The clarification data is divided into 3 parts: training set (14772 samples), development set (500 samples), and test set 1 (500 samples). Real miss-recognized utterances are used as test set 2 (208 samples, partial errors only). A 3-gram language model is trained with the SRILM toolkit for all generation models by using all CQs of the training data and a Chinese spoken language corpus1 (401,772 utterances). Two decoders are used in experiments. The baseline decoder is following the conventional method using only initial phrase table, and the other one is our two-step decoder using both generalized and un-generalized phrase tables.

A. Evaluation methodology

Previous works are all use fixed clarification questions or templates to prompt a clarification dialogue, so it is impossible

to compare our results to them. BLEU score is widely used to evaluate the quality of machine translation and is used in this work as an automatic metric of evaluating the quality of generated CQs. However, only BLEU metric is not sufficient. There may be many ways to clarify an miss-recognized utterance, all of these CQs are reasonable. BLEU score can only reflects to what extent the output is close to the input, but cannot reflect the reasonableness of the output. In order to analysis the rationality of outputs, in our experiments, human evaluation method is also adopted. All generated CQs are manually judged and divided into 3 classes: (1) Perfect, the generated CQ is reasonable and meets the oral expression style; (2) Acceptable, the generated CQ is reasonable but not fluent in expression; (3) Unacceptable, the generated CQ is not reasonable. The proportion of each category can well reflect the system performance.

B. Experiments on CQ Generation

We compare the size of different phrase tables, as shown in Table III. Through only 14.57% of initial phrases can be generalized, their generalization ability is amazing. The size of phrase table extracted by using our method is about 4 times as much as that of the initial phrase table.

TABLE III. SIZE OF EXTRACTED PHRASE TABLES

Phrase Table Type		**Size**
Initial phrase		254.7k
Generalized phrase	before generalization	37.1k
	after generalization	883.3k
Non-generalized phrase		217.4k
Generalized + Non-generalized phrase		1101k

Experimental results on simulated data (Test set 1) are shown in Table IV. Totally three models are trained based on different phrase tables. The first model is trained on initial phrases. Experimental results show that 28.4% of outputs generated by the model are judged as unacceptable. In comparison, the third model trained on both un-generalized and generalized phrases only has a 14.8% unacceptable rate. It indicates a good generalization ability of our method. The second model trained on only generalized phrases shows a counterpart performance with the first model. However, we should notice that all of generalized phrases are extracted from only 37.1k initial phrases, so the result can also reflects the effectiveness of our method. BLEU score of three models can also prove that the generalized phrase can help improve the quality of CQ generation.

TABLE IV. SIZE OF EXTRACTED PHRASE TABLES

Model	**BLEU-3**	**Human Evaluation**		
		Perfect	Acceptable	Unacceptable
Initial phrase only	28.7	55.4%	16.2%	28.4%
Generalized phrase only	30.3	53.2%	20.6%	26.2%
Generalized + un-generalized	**37.6**	68.8%	16.4%	**14.8%**

TABLE V. SIZE OF EXTRACTED PHRASE TABLES

Model	**Human Evaluation**		
	Perfect	Acceptable	Unacceptable
Initial	57.6%	20.3%	22.1%
Gen.+ Non-Gen.	70.7%	16.3%	**13.0%**

Experimental results on real data (Test set 2) are shown in Table V. In comparison with experimental results on simulated data, both models show a better performance. This is because errors in simulated data are randomly selected and are more complicated than errors of a real speech recognition system. Experimental results indicate that our method can be well applied to real speech recognition systems.

V. CONCLUSIONS

We propose a new approach to CQ generation for prompting clarification dialogue, which can be used to handling speech recognition errors. Our method can be seen as an imitation of clarification behavior in human-human conversation. The problem is seen as a monolingual SMT task that translates miss-recognized utterance into a CQ. A PB-SMT framework is used to generate CQs and proved to be feasible. In order to adapt to CQ generation task and improve system performance, a new method of extracting generalized phrase pair is proposed. Corresponding to it, a two-step decoder is used for CQ generation. Experiments are carried out on both simulated data and real data, experimental results show that our method can generate reasonable CQs, and generalized phrase table combined with two-step decoder achieves better performance than that of conventional PB-SMT system. The work provides a general way of prompting clarification dialog according to speech recognition errors, and can be applied to various speech interface applications.

ACKNOWLEDGMENT

This research is supported by Science Foundation of BLCU (supported by “the Fundamental Research Funds for the Central Universities”) (Approval No. 11JBB037, 12YBG02).

REFERENCES

[1] Schlangen D., “Causes and Strategies for Requesting Clarification in Dialogue,” In Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue. Boston, 2004, pp. 136-143.

[2] Purver M., “CLARIE: Handling Clarification Requests in a Dialogue System,” Research on Language & Computation, vol.4, no. 2, pp. 259-288, 2006.

[3] Skantze G., “Exploring human error recovery strategies: Implications for spoken dialogue systems.”, Speech Communication, vol. 45, no. 3, 2005, pp. 325-341.

[4] Lewis C., Fabbrizio G.D., “A clarification algorithm for spoken dialogue system,” In Proceedings of. ICASSP2005, Philadelphia, 2005.

[5] Misu, T., Kawahara, T. “Dialogue strategy to clarify user’s queries for document retrieval system with speech interface”. Speech Communication, vol. 48, no. 9, 2006, pp. 1137-1150.

[6] Sangkeun J., Cheongjae L., Gary G. L., “Three Phase Verification for Spoken Dialog Clarification,” In Proceedings of IUI2006, Sydney, 2006, pp. 55-61.

[7] Quirk, C., Brockett, C., Dolan, W. “Monolingual machine translation for paraphrase generation.” In Proceedings of EMNLP. pp.142-149, 2004.

[8] Koehn P., Och F. J., Marcu D., “Statistical phrase-based translation,” In Proceedings of NAACL/ HLT, 2003, pp. 48-54.

[9] Gao S., Xu B., Huang T. Y., “A new framework for Mandarin LVCSR base on one-pass decoder,” In Proceedings of ISCSLP 2000, Beijing, 2000, pp. 49-52.

[10] Stuttle M. N., Williams J. D., Young S. “A Framework for Dialogue Data Collection with a Simulated ASR Channel.”, In Proceedings of ICSLP, Jeju, South Korea, 2004.

[11] Wessel F., Schluter R., Macherey K., Ney H., “Confidence measures for large vocabulary continuous speech recognition”, IEEE Transactions on Speech and Audio Processing, vol. 9, no. 3, 2001, pp. 288-298.

Published by Atlantis Press, Paris, France.