Cross-lingual Dialog Model for Speech to Speech Translation

Emil Ettelaie, Panayiotis G. Georgiou, Shrikanth Narayanan

Speech Analysis and Interpretation Lab
Department of Electrical Engineering - Systems
Viterbi School of Engineering
University of Southern California
ettelaie@usc.edu

Abstract

Speech understanding through concept classification offers a possible way of machine translation in speech-to-speech translation systems and can be used in conjunction with conventional statistical machine translation. While correct concept classification offers the promise of obtaining well-formed target language speech output, the approach does not scale well to large number of concepts. Importantly, it is also critical to know when to accept or reject the classifier. We formulate the speech classification as a MAP estimation problem to derive the understanding model and improve its performance by incorporating dialog context information. Specifically, for a two-way speech translation system, a classification scheme is derived here that utilizes context information from both sides of the conversation through an n-gram dialog model. The method was evaluated using data from an English-Farsi trans-lingual doctor-patient dialog system and its classification and rejection accuracies were compared to those of a baseline system with an understanding model only. The benefit of incorporating context with the proposed dialog model provided a modest improvement in classification accuracy (about 5% relative error reduction) and a significant improvement in the rejection accuracy (up to 31.4% relative reduction in error).

Index Terms: speech-to-speech translation, machine translation, speech understanding, concept classification, n-gram models, dialog modeling

1. Introduction

The primary goal of speech to speech (S2S) translation systems is to mediate communication between two people that do not share the same language. Hence, it becomes critical to assure accurate concept transfer beyond merely maximizing speech recognition accuracy and machine translation scores. Different machine translation (MT) approaches have been adopted in current speech-to-speech (S2S) systems. Translation engines developed as statistical machine translators [10] or utterance classifiers based on understanding [1], [12] or inter-lingua [11], have been integrated into S2S systems. While the methods based on utterance classification can yield high quality of translated speech within their domain, they suffer accuracy degradation as the number of classes covering the domain increases. Besides, they do not scale well to handle multiple concepts per utterance. In contrast, the performance degradation is more gradual for the statistical MT methods although the quality of their output speech is generally inferior to what classifiers produce due to the potential lexical and syntax errors. A combination of these methods is a popular way of addressing the limitations of each [1], [2], [8]. The idea would be to use the classifier in cases when it works, and fall back to the SMT otherwise. One challenge in doing so is to figure out how to choose amongst, and rank, the options provided by the different schemes. In this work, we focus on designing a system that aims to combine effectively a classifier with a traditional phrase-based statistical machine translator (SMT). Specifically, in designing such a system the following issues need to be considered: 1) improving the classifier accuracy, and 2) enabling a preference mechanism that helps the system make robust selections between the two translation methods during an S2S interaction.

This raises the motivation to seek ways of using additional sources of information, beside the text from the recognized speech, to improve the classification performance and to enhance the ability to reject low confidence classification results in favor of the statistical MT output. In an attempt to achieve those goals, a method of using cross-lingual dialog information in conjunction with a method for rejecting low confidence classifier output is proposed. In such systems, utterances from each side of conversation are statistically mapped to predefined concepts. For each side, these mappings carry information about the potential concept of choice in the other side, and are tracked and exploited during classification. This sort of dialog information has been used previously in different applications to enhance the performance of speech utterance classification. For instance, in [3] the classification problem has been formulated in a way that the use of dialog information increases the accuracy of the classifier in a categorical classification task. A similar MAP classification approach is used here to utilize the information that is carried by concept history sequence from both sides of the conversation.

Taking the MAP classification formulation leads to a practical framework that converts the problem into two straightforward modeling problems. The first model, i.e. understanding model, presents the statistical relation between the transcribed utterances and the concepts. This type of modeling has been used for other applications in [5] and [6]. The second model is the statistical dialog model that statistically connects the concepts expressed by both sides of the conversation. The proposed method is evaluated with a corpus of doctor-patient English-Farsi dialogs [1, 7].

The organization of this paper is as follows: In section 2, concept classification approach for speech translation is reviewed and the understanding model is described. Section 3, starts with the formulation of the problem for the S2S mediation context, and continues with the derivation of a method that uses the dialog model. The system and the experiment set up are explained in section 4 and results were discussed in section 5.

2. Concept Classification and Understanding Model

Using concept classifiers for speech translation purpose has been investigated based on covering the target dialog domain by several concept classes. For example, in the medical domain dialogs of [1, 2], the doctor side of the conversation was mapped into 1200 pre-specified classes and the patient side with around 400 classes. Associated with each class are representative surface form instantiations that convey the concept in the target language in the best way. The classifier then tries to map the source language spoken utterances into one of the predefined concepts.

Upon a successful mapping the representing phrase of the wining concept class is played out in the target language.

If the dialog domain is partitioned in a set of concept classes $\aleph = \{C^{(1)}, C^{(2)}, ..., C^{(|\aleph|)}\}$ the classification task can be formulated as the following maximum a posteriori estimation.

$$\hat{C}_t = \arg\max_{C} P(C \mid O_t) \quad (1)$$

where $C \in \aleph$ and O_t is the acoustic observation of the spoken utterance in the source language at turn t and $\hat{C}_t$ is the estimated concept of that utterance.

In practice, the above estimation is implemented in two well-known steps. First, the acoustic signal is transcribed by an automatic speech recognizer (ASR) and then a text classifier maps the transcription to a concept by using its words as classification features. With no prior knowledge about the concepts this would reduce to a maximum likelihood classifier, i.e.,

$$\hat{C}_t = \arg\max_{C} P(\hat{\mathbf{W}}_t \mid C) \quad (2)$$

Here, $\hat{\mathbf{W}}_t$ is the vector of words generated by ASR as the transcription of O_t. The likelihood function $P(\hat{\mathbf{W}}_t \mid C)$ can be approximated by a language model (LM) built specifically for each $C \in \aleph$. All of these concept-specific language models form what is known as the "Understanding Model" [3]. In section 4 building the understanding model is explained in detail.

A two way speech to speech translation system needs two classifiers similar to Eq. (2). Each one of those should have an understanding model in the language of its corresponding side. However, depending on the application the concept sets can be different for each one, which will make the system asymmetric. An example of such a system is explained in [1] and [2]. It facilitates doctor-patient dialogs in which one side (doctor) often drives the conversation by asking questions.

3. Dialog Model

The system described in section 2 has the drawback of not utilizing any dialog context information. Especially in an asymmetric two way system where the driver side's utterances (such as the doctor who controls the dialog flow in a typical doctor-patient interactions) seem to carry some (if not much) information about what the other interlocutor says. An approach to incorporating such information is presented in [3] for a human-machine dialog system. A similar method can be used to incorporate dialog information in a two way concept classification task.

Let us consider a two way mediation system where each interlocutor has his own concept set (such as in a Q- task). Such an asymmetric system will have a concept set $\Re = \{R^{(1)}, R^{(2)}, ..., R^{(|\Re|)}\}$ for the driving side (side A) and a different set $\aleph = \{C^{(1)}, C^{(2)}, ..., C^{(|\aleph|)}\}$ for the other side (B). The goal is to come up with a new classifier for side B, that uses the decisions made by side A, with the hope that the extra information would lead to a better classification accuracy for side B.

With the assumption that the chain of dependency is limited to only one cycle of conversation (note that a first order Markov model for dialog history was found to be effective in [3]), the classifier for side B can be rewritten as the following a maximum a posterior estimator.

$$\hat{C}_t = \arg\max_{C} P(C \mid O_t, R_t) \quad (3)$$

where $C \in \aleph$, $R_t \in \Re$, and O_t is the acoustic observation of side B in conversation turn t. Since in practice the transcription of O_t is needed for classification, we reformulate (3) as the following maximization.

$$\max_{C,\mathbf{W}} P(C, \mathbf{W} \mid O_t, R_t) \quad (4)$$

$\mathbf{W}$ is a potential transcription of O_t. In the absence of any prior information of R_t the above maximization is equivalent to

$$\max_{C,\mathbf{W}} P(O_t \mid \mathbf{W}, R_t, C) \cdot P(\mathbf{W}) \cdot P(\mathbf{W} \mid C, R_t) \cdot P(C \mid R_t) \quad (5)$$

Two assumptions are made here to make the maximization problem practically feasible:

1. The first assumption is $P(O_t \mid \mathbf{W}, R_t, C) \approx P(O_t \mid \mathbf{W})$ which means that the acoustic observations do not depend on any side's chosen concepts [3].
2. It is also assumed that the transcription of side B's utterance and side A's concept are independent, i.e., $P(\mathbf{W} \mid C, R_t) = P(\mathbf{W} \mid C)$.

These assumptions help split Eq. (5) as the following two step maximization.

$$\hat{\mathbf{W}}_\mathbf{t} = \arg\max_{\mathbf{W}} P(O_t \mid \mathbf{W}) \cdot P(\mathbf{W}) \quad (6)$$

$$\hat{C}_t = \arg\max_{C} P(\hat{\mathbf{W}}_\mathbf{t} \mid C) \cdot P(C \mid R_t) \quad (7)$$

This decomposition is greatly beneficial from practical point of view. According to (6) $\hat{\mathbf{W}}_\mathbf{t}$ can be the output of an ASR with acoustic and language models that estimate $P(O_t \mid \mathbf{W})$ and $P(\mathbf{W})$ respectively. Eq. (7) shows that the concept of the utterance spoken by side B can be estimated using the ASR output and the concept chosen by side A. While estimator (2) only uses the information from one side, (7) uses the statistical dependency of concepts in both sides of the dialog, i.e., $P(C \mid R_t)$. This dependency can be estimated by dialog model.

Hence, in practice, the Eq. (7) is used for concept estimation as

$$\hat{C}_t = \arg\max_{C} P_U(\hat{\mathbf{W}}_\mathbf{t} \mid C) \cdot \left[P_D(C \mid R_t)\right]^{\gamma} \quad (8)$$

where P_U and P_D are the understanding model and the dialog model respectively. These functions are the imperfect estimates of probabilities in Eq. (7). The exponential weight γ is introduced here to emphasize (or deemphasize) the effect of the dialog model. Eq. (8) is in fact a log-linear combination of the scores from two models. Section 4 contains the detail of tuning γ to reach the best performance.

4. Task Description

An S2S MT system for English-Farsi interactions in the medical domain was chosen as the framework to test the new modeling of Eq. (8). The system is a two way S2S system that facilitates the medical interviews [1], [2], [8]. The languages are English and Farsi for the doctor and patient side, respectively. Beside a statistical MT engine, the system also uses concept classification as an alternative translation method (Figure 1). The system is asymmetric in the sense that concept sets for the sides are not similar. The concepts were manually assigned using expert resources. In the type of conversations that the system is designed for in [1, 2], the doctor predominantly controls the flow of the dialog.

While the acoustic and language models of Eq. (6) are embedded in the system's two ASRs [9], the focus of this task was to build understanding and dialog models and compare the system performance under two conditions with and without the dialog model.

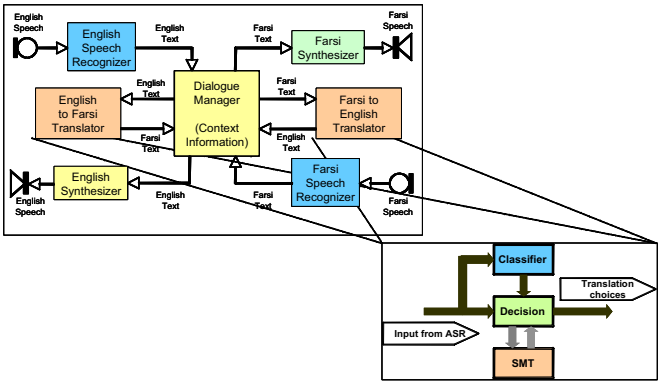

Figure 1: *The English-Farsi S2S system with a dual translation scheme involving a classifier and an SMT*

2.1. Categorical Understanding Model: Concept Classification

The understanding model consists of the LMs built for each concept-class. The data needed to build a LM for a specific concept are in the form of paraphrases that convey that concept. Such data were collected in different ways. For instance, to collect paraphrases for the doctor side, a website was set up and several people were invited to enter paraphrases for each concept. Data were also collected from standardized patient based methods [7]. Data were also collected directly from a few small groups of Farsi speakers for each concept class in the patient side. Table 1 shows the collected data statistics.

For each class a trigram LM was built using the SRILM toolkit with Ristad's natural discounting law [4]. Transcriptions from in-domain conversations were also used to build a background LM. Each class-specific LM was interpolated with that background model to create a smoother model for each class [4]. That sort of smoothing was used in the task reported in [5] and a similar method was introduced in [6]. The interpolation weight λ of the two models requires optimization through a development set representative of the usage conditions.

In addition to the concept classes, a "rejection" or "null" class was also included in the concept set for each side. That class represents the utterances that do not convey any of the covered concepts and therefore should be rejected. The LM of the null class was only trained on the background data. Presence of a null class shows a great advantage when the concept set does not cover whole the dialog domain. From a practical point of view, this was especially critical because in such an unrestricted conversation scenario, at best, the classifier is designed for, and can capture, only a fraction of the concepts conveyed.

For each side of the dialog an understanding model was built using the class-specific LM. Manually transcribed and translated data of human-human, monolingual (English), non-mediated in-domain interactions were annotated and used as a development set. Thus the interpolation weight λ of the understanding model was not optimized for the case of noisy ASR outputs. The data is described in Table 2.

2.2. Dialog Model

In this task, the dialog model represents the statistical dependency of the patient's concept on the concept expressed by doctor's utterance in the same conversation cycle. For training the model, the manual transcription of audio data recorded from doctor-patient conversations were used [7]. A set of 15,411 conversation turns (i.e., doctor's utterance followed by patient's response) were selected for unsupervised training of the dialog model. Each utterance pair was first converted to a pair of concept tags using the classifier of the corresponding side. The tag pairs were then used to train a bigram dialog model using SRILM with no discounting scheme.

Table 1: Training data for understanding model

Specification	Doctor	Patient
Language	English	Farsi
Type	Text	Text
Number of concept classes	1,269	364
Training data for class-specific LMs	9,894 lines (60,050 words)	27,459 lines (182,751 word)
Training data for background LM	25,305 lines (224,642 words)	42,870 (263,683 words)

2.3. Combining Understanding and Dialog Models

The concept classifier makes decisions only based on its understanding model. The goal of this task was to improve the accuracy of the decisions in the patient side by using the information from doctor's decisions through the dialog model.

Table 2: Development and testing data (Both data sets were manually tagged for performance measurement.)

Application	Side	Language	Type	Size of the data [utterance]
Set A	Doctor	English	Manual transcription	106
	Patient	Farsi	Human translation	106
Set B	Doctor	English	Manual transcription	252
	Patient	Farsi	ASR output	252

To apply the dialog model, the scores generated by classifier must be combined with the dialog model scores and the overall scores should be used to select the right class. However, the effect of dialog model can be emphasized (or deemphasized) by the introduction of the exponential weight in Eq. (8). To set this parameter to an optimal value a development data set was prepared (Table 2). The development data was manually annotated for performance measurement. A one-dimensional search for the parameter led to a setting that gave the minimum classification error on the development data. That setup was then frozen and used to test the system.

5. Results

To validate the benefits of applying a dialog model, two sets of experiments were conducted:

1. Set A, as described in Table 2 was used for development, and Set B as a test set. As a result there is a mismatch of training and testing conditions
2. Sets A and B are combined and using the leave one out technique about two thirds of the data were used for development and the remaining for testing and the test is repeated three times

In Table 2 the patient side of Set A was manually translated and transcribed, while Set B was generated by running a Farsi ASR [9] on the recorded audio files.

The classifier was evaluated on these transcripts and compared with human class annotations. The performance of the

Table 3: *Classifier accuracy with and without dialog model for Experiment 1*

Experiment 1	Development	Testing
All Data		
Baseline (w/o dialog model)	50.00%	41.30%
w/ dialog model	53.80%	44.80%
Relative Error Reduction	**7.60%**	**5.96%**
Data annotated as "null" (rejection)		
Baseline	41.90%	48.10%
w/ dialog model	54.80%	64.40%
Relative Error Reduction	**22.20%**	**31.41%**
Excluding the data recognized as "null"		
Baseline	51.30%	33.30%
w/ dialog model	54.80%	35.40%
Relative Error Reduction	**7.19%**	**3.15%**

Table 4: *Classifier accuracy with and without dialog model for Experiment 2*

Experiment 2	Development	Testing
All Data		
Baseline (w/o dialog model)	56.80%	58.37%
w/ dialog model	58.62%	60.37%
Relative Error Reduction	**4.21%**	**4.80%**
Data annotated as "null" (rejection)		
Baseline	90.00%	89.55%
w/ dialog model	91.37%	91.13%
Relative Error Reduction	**13.74%**	**15.07%**
Excluding the data recognized as "null"		
Baseline	55.52%	55.60%
w/ dialog model	59.91%	60.18%
Relative Error Reduction	**9.87%**	**10.31%**

classifier without the dialog model was acquired as the baseline. Then the classifier with the dialog model was applied on the same data.

Table 3, presents the results of experiment 1, where there is a mismatch of development and test conditions in the optimization of the understanding model and shows the accuracy obtained by the patient side classifier before and after engaging the dialog model. According to these results, a relative decrease of 5.96% in the classification error was achieved by applying the dialog model. The improvement is due to a more elaborate modeling based on of Eq. (3) that utilizes some of the dialog information. Table 3 also shows a significant improvement (relative reduction of 22.2% for development and 31.41% for testing data) in the rejection accuracy achieved by applying the dialog model.

However, looking at the overall improvement by itself can be misleading since it could merely be due to a higher score given to the null concept by dialog model. Therefore, the accuracy was also measured without the utterances that were classified as the null concept. Since the accuracy improvement was also observed in this case (3.15%), we can deduct that the dialog model has helped the classification of non-null utterances in addition to improving the rejection accuracy (31.41%).

The results of experiment 2 are shown in Table 4 where we again observe a 4.80% relative error reduction. The relative improvement in rejection accuracy is smaller, but still significant (15.07%) than in experiment 1, likely due to the optimization of the understanding model, which results in smaller margins of potential improvement.

6. Conclusion

The formulation of the concept classification as a maximum a posteriori estimation was adopted and extended in a practical way that was suitable for the S2S mediation scenarios. The formulation decomposes the estimation task into two separate well-known steps of speech recognition and text classification with four familiar components: Acoustic Model, Language Model, Understanding Model, and Dialog Model. The resultant classification scheme not only is based on an understanding model but also uses a dialog model. That provides a way to incorporate dialog information directly into the speech translation task.

For a two-way S2S MT system that is based on the concept classification, the deployment of the dialog model opens up a way to use the information from the both sides of conversation. For the English-Farsi doctor patient dialog system used in this work, a relative reduction in error of 4.80-5.96% in the classifier performance indicates that using the information from one side of the conversation (driving side, in this task) can improve the quality of the translations of the other interlocutor's speech.

More notably, applying the dialog model also improved the accuracy of the rejection significantly (15.07-31.41% relative error reduction). This leads to a more accurate detection of the cases that classifier fails to produce a reliable result and therefore indicating that the fallback selection of the output from the statistical MT will be more appropriate.

Our future work focuses on further improving the classifier accuracy by incorporating rich speech information such as conveyed by prosody, and by incorporating acoustic and lexical confidence scores directly within the classifier model.

7. Acknowledgements

This work was supported by the DARPA Babylon/CAST program, contract N66001-02-C-6023.

8. References

[1] R. Belvin et al, "Transonics: A practical speech-to-speech translator for English-Farsi medical dialogs," *In Proc. of The Association of Computational Linguistics*, Univ. of Michigan, Ann Arbor, June 2005.

[2] S. Narayanan, et al., "Transonics: A speech to speech system for English-Persian interactions," *IEEE Automatic Speech Recognition and Understanding Workshop*, ASRU, 2003.

[3] A. Potamianos, S. Narayanan, and G. Riccardi, "Adaptive categorical understanding for spoken dialog systems," *IEEE Trans. Speech and Audio Processing*, 13(3):321-329, 2005.

[4] A. Stolcke, "SRILM -- An extensible language modeling toolkit," *In Proc. Intl. Conf. on Spoken Language Processing*, vol. 2, pp. 901-904, Denver, 2002.

[5] C. Chelba, M. Mahajan, and A. Acero, "Speech utterance classification," *In Proc. of the Int. Conf. on Acoustics, Speech, and Signal Processing*. Hong Kong, Apr, 2003.

[6] A. Sethy, P. G. Georgiou, and S. Narayanan, "Building topic specific language models from webdata using competitive models," *In Proc. of Eurospeech*, 2005.

[7] R. Belvin, W. May, S. Narayanan, P. Georgiou, and S. Ganjavi, "Creation of a doctor-patient dialog corpus using standardized patients," *In Proc. LREC*, Lisbon, Portugal, 2004.

[8] S. Narayanan, et al., "The Transonics spoken dialog translator: an aid for English-Persian doctor-patient interviews," *In Proc. of AAAI Fall Symposium*, 2004.

[9] N. Srinivasamurthy and S. Narayanan, "Language-adaptive Persian speech recognition," *In Proc. Eurospeech*, Geneva, Switzerland, 2003.

[10] H. Ney, et al., "Algorithms for statistical translation of spoken language,". *IEEE Transactions on Speech and Audio Processing*. 8(1):24-36, January 2000.

[11] A. Waibel, et al., "Speechalator: two-way speech-to-speech translation on a consumer PDA," *In Proc. Eurospeech*, Geneva, Switzerland, 2003.

[12] B. Zhou, D. Dechelotte, and Y. Gao, "Two-way speech-to-speech translation on handheld devices," *In Proc. Int. Conf. of Spoken Language Processing*, Korea, Oct. 2004.