INCORPORATING DISCOURSE CONTEXT IN SPOKEN LANGUAGE TRANSLATION THROUGH DIALOG ACTS

Vivek Kumar Rangarajan Sridhar, Shrikanth Narayanan

Speech Analysis and Interpretation Laboratory
University of Southern California
Viterbi School of Electrical Engineering
vrangara@usc.edu, shri@sipi.usc.edu

Srinivas Bangalore

AT&T Research Labs
180 Park Avenue
Florham Park, NJ 07932, U.S.A.
srini@research.att.com

ABSTRACT

Current statistical speech translation approaches predominantly rely on just text transcripts and are limited in their use of rich contextual information such as prosody and discourse function. In this paper, we explore the role of discourse context characterized through *dialog acts* (DAs) in statistical translation. We present a bag-of-words (BOW) model that exploits DA tags in translation and contrast it with a phrase table interpolation approach presented in previous work. In addition to producing interpretable DA-annotated target language translations through our framework, we also obtain consistent improvements in terms of automatic evaluation metrics such as lexical selection accuracy and BLEU score using both the models. We also analyze the performance improvements per DA tag. Our experiments indicate that *questions*, *acknowledgments*, *agreements* and *appreciations* contribute to more improvement in comparison to *statements*.

Index Terms— Enriched translation, dialog acts, bag-of-words model

1. INTRODUCTION

While machine processing of speech has advanced significantly, it is still largely compartmentalized. For instance, automatic speech recognition typically deals with orthographic transcription of the speech and hence is insufficient for capturing the context beyond words. Enriched transcription has emerged as a unifying theme in spoken language processing, combining automatic speech recognition and natural language processing with the goal of producing richly annotated speech transcriptions. In this paper, we investigate the use of rich annotation in the form of dialog act tags in spoken language translation. The proposed framework captures shallow discourse structure of the source text using an automatic dialog act tagger and exploits the detected annotation within the translation process.

Dialog act tags have been previously used in the VERBMOBIL statistical speech-to-speech translation system [1]. In that work, the predicted DA tags were mainly used to improve speech recognition, semantic evaluation, and information extraction modules. A dialog act-based translation module in VERBMOBIL was presented in [2]. The module was mainly designed to provide robustness in the translation process in case of defective input from the speech recognition system. Ney et al. [3] proposed a statistical translation framework to facilitate the translation of spoken dialogues in the VERBMOBIL project. Their framework was integrated into the VERBMOBIL prototype system along with the dialog act-based approach developed in [2]. Discourse information in the form of speech acts has also been used in interlingua based translation systems [4] to map input text to semantic concepts, which are then translated to target text.

Our objective in this work is two-fold. First, we are interested in capturing source language discourse information in terms of dialog acts and exploiting them within the translation process to improve overall translation quality. Second, our scheme augments conventional speech translation with rich annotations that can aid in disambiguation and improved interpretation of translation hypotheses, thus, enabling better cross-lingual dialog in two-way translation devices.

2. ENRICHED TRANSLATION USING DAs

In this section, we formulate the problem of using rich annotations in speech translation. If S_s, T_s and S_t, T_t are the speech signal and equivalent textual transcription in the source and target language, L_s the enriched representation for the source speech, we can formalize our proposed S2S translation as shown in Figure 1. Eq.(3) is obtained from Eq.(2) through conditional independence assumptions. Even though the recognition and translation can be performed jointly [5], typical S2S translation frameworks compartmentalize the ASR, MT and TTS with each component maximized for performance individually. T_s^*, T_t^* and S_t^* are the arguments maximizing the ASR, MT and TTS components respectively. L_s^* is the rich annotation detected from the source speech signal and text, S_s and T_s^* respectively. In this work, we do not address the speech synthesis part and assume that we have access to the reference transcripts or 1-best recognition hypothesis of the source utterances. The rich annotations (L_s) can be syntactic or semantic concepts [6], prosody [7, 8], or, as in this work, dialog act tags.

2.1. Automatic dialog act tagging

In this work, we use a dialog act tagger trained on the Switchboard DAMSL corpus [9] using a maximum entropy (maxent) model. The original tagset of 375 unique tags was clustered to obtain 42 dialog tags as in [9]. In addition, we also grouped the 42 tags into 7 disjoint classes, based on the grouping presented in [10]. The simplified tagset consisted of the following 7 classes: *statement, acknowledgment, abandoned, agreement, question, appreciation, other*. Detailed explanation of the maxent dialog act tagger can be found in [11]. Table 1 summarizes the dialog act tagging performance on a test set comprising 29K sentences from the SWBD-DAMSL corpus.

The detected dialog act tags (L_s^*) can be exploited in statistical machine translation in a variety of ways. In the following section,

$$S_t^* = \arg\max_{S_t} P(S_t|S_s) \tag{1}$$

$$\begin{aligned} P(S_t|S_s) &= \sum_{T_t,T_s,L_s} P(S_t,T_t,T_s,L_s|S_s) \\ &= \sum_{T_t,T_s,L_s} P(S_t|T_t,T_s,L_s,S_s).P(T_t|T_s,L_s,S_s).P(L_s|T_s,S_s).P(T_s|S_s) \end{aligned} \tag{2}$$

$$\approx \sum_{T_t,T_s,L_s} P(S_t|T_t,L_s).P(T_t|T_s,L_s).P(L_s|T_s,S_s).P(T_s|S_s) \tag{3}$$

$$\max_{S_t} P(S_t|S_s) \approx \max_{S_t} P(S_t|T_t^*,L_s^*).\max_{T_t} P(T_t|T_s^*,L_s^*).\max_{L_s} P(L_s|T_s^*,S_s).\max_{T_s} P(T_s|S_s) \tag{4}$$

Augmented Text-to-Speech **Enriched Machine Translation** **Rich Annotation** **Speech Recognition**

Fig. 1: Formulation of the proposed enriched speech-to-speech translation framework

Cues used (current utterance)	Accuracy (%)	
	42 tags	**7 tags**
Lexical	69.7	81.9
Lexical+Syntactic	70.0	82.4
Lexical+Syntactic+Prosodic	70.4	82.9

Table 1: Dialog act tagging accuracies for various cues on the SWBD-DAMSL corpus.

we propose a bag-of-words approach for exploiting dialog act tags in translation. Furthermore, we compare the proposed approach with a phrase table interpolation scheme for utilizing dialog act tags previously presented in [12].

2.2. Exploiting DAs in a bag-of-words translation model

Conventional phrase-based translation relies on learning phrasal associations that are derived from word alignment information. The target bag-of-phrases is typically reordered using a target language model. As a result, there is little emphasis on global lexical reordering which may be necessary for certain language pairs. In contrast, a bag-of-words approach to translation estimates the probability of each target word independently in the context of the entire source sentence. The detected bag-of-words can then be reordered using a language model. Such a bag-of-words (BOW) approach to translation was first presented in [13] and is illustrated in Figure 2.

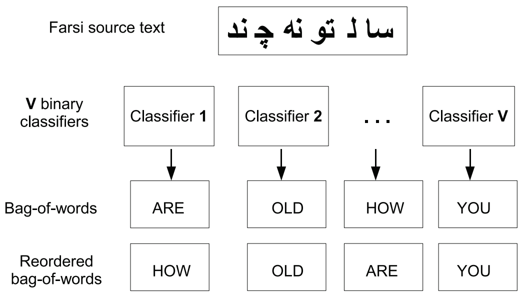

Fig. 2: Illustration of bag-of-words approach to translation

In this work, we extend the bag-of-words approach by exploiting dialog act tags and thus, enriching translation. We treat the target sentence as a BOWs assigned to the source sentence and it's corresponding dialog act tag. The objective here is, given a source sentence and the dialog act tag, estimate the probability of finding a given word in the target sentence. Since, each word in the target vocabulary is detected independently, one can use simple binary static classifiers. The classifier is trained with word n-grams and dialog act obtained from the source sentence T_s $((BOWgrams(T_s), L_s)$. During decoding, the words with conditional probability greater than a threshold θ are considered as the result of lexical choice decoding (Eq.(5)). We use a binary maximum entropy technique with L1-regularization for training the bag-of-words lexical choice model. The machine learning toolkit LLAMA [14] was used for training the maxent model.

$$BOW^*_{T_t} = \{T_t|P(T_t|BOWgrams(T_s), L_s) > \theta\} \tag{5}$$

For reconstructing the correct order of words in the target sentence, we consider all permutations of words in $BOW^*_{T_t}$ and rank them using a target language model. In this work, we used a separate language model for each dialog act, created by interpolating the DA-specific language model with the baseline language model obtained from the entire data. We control the length of the target sentences by varying the parameter θ.

2.3. Comparison with phrase-based translation

We have previously proposed a phrase table interpolation scheme for exploiting dialog act tags in phrase-based translation [12]. To summarize the scheme briefly, to each phrase translation table belonging to a particular DA-specific translation model, we append those entries from the baseline model that are not present in the phrase table of the DA-specific translation model. The appended entries are weighted by a factor α.

$$\begin{aligned} (T_s \to T_t)_{L_s^*} = (T_s \to T_t)_{L_s} \cup \{\alpha.(T_s \to T_t) \\ s.t.\ (T_s \to T_t) \notin (T_s \to T_t)_{L_s}\} \end{aligned} \tag{6}$$

where $(T_s \to T_t)$ is a short-hand[1] notation for a phrase translation table. $(T_s \to T_t)_{L_s}$ is the DA-specific phrase translation table, $(T_s \to T_t)$ is the phrase translation table constructed from the entire

[1] $(T_s \to T_t)$ represents the mapping between source alphabet sequences to target alphabet sequences, where every pair $(t_1^s, \cdots, t_n^s, t_1^t, \cdots, t_m^t)$ has a weight sequence $\lambda_1, \cdots, \lambda_5$ (five weights).

	Training						Test					
	Farsi	Eng	Jap	Eng	Chinese	Eng	Farsi	Eng	Jap	Eng	Chinese	Eng
Sentences	8066		12239		46311		925		604		506	
Running words	76321	86756	64096	77959	351060	376615	5442	6073	4619	6028	3826	3897
Vocabulary	6140	3908	4271	2079	11178	11232	1487	1103	926	567	931	898
Singletons	2819	1508	2749	1156	4348	4866	903	573	638	316	600	931

Table 2: Statistics of the training and test data used in the experiments.

		F-score (%)			BLEU (%)		
		w/o DA tags	w/ DA tags		w/o DA tags	w/ DA tags	
Framework	Language pair		7tags	42tags		7tags	42tags
BOW model	Farsi-Eng	58.00	59.14	59.35	15.95	16.99	17.12
	Japanese-Eng	79.50	79.82	79.93	42.54	44.70	44.98
	Chinese-Eng	68.83	69.70	69.91	54.76	55.98	56.14
Phrase-based translation [12]	Farsi-Eng	56.46	57.32	57.74	22.90	23.50	23.75
	Japanese-Eng	79.05	79.40	79.51	54.15	54.21	54.32
	Chinese-Eng	65.85	67.24	67.49	48.59	52.12	53.04

Table 3: F-measure and BLEU scores for the two different translation schemes with and without use of dialog act tags.

data and $(T_s \rightarrow T_t)_{L_s^*}$ is the newly interpolated phrase translation table. The interpolation factor α is used to weight each of the four translation scores (phrase translation and lexical probabilities for the bilanguage) with the phrase penalty remaining a constant.

3. DATA

We report experiments on three different parallel corpora: Farsi-English, Japanese-English and Chinese-English. The Farsi-English data used in this paper was collected for doctor-patient mediated interactions in which an English speaking doctor interacts with a Persian speaking patient [15]. The Japanese-English parallel corpus is a part of the "How May I Help You" (HMIHY) [16] corpus of operator-customer conversations related to telephone services. The Chinese-English corpus corresponds to the IWSLT06 training and development set [17] . The data are traveler task expressions. Table 2 presents statistics of the corpora used in our experiments.

4. EXPERIMENTAL RESULTS

The lexical selection accuracy and BLEU scores for the three parallel corpora using the two translation schemes described in Section 2.2 and 2.3 are presented in Table 3. Lexical selection accuracy is measured in terms of the F-measure derived from recall ($\frac{|Res \cap Ref|}{|Ref|} * 100$) and precision ($\frac{|Res \cap Ref|}{|Res|} * 100$), where Ref is the set of words in the reference translation and Res is the set of words in the translation output. For both the statistical translation frameworks, adding dialog act tags (either 7 or 42 tag vocabulary) consistently improves both the lexical selection accuracy and BLEU score for all the language pairs. While the BOW model provides higher lexical selection accuracy, the phrase-based translation provides better BLEU score. In the BOW model, we detect each word in the target vocabulary independently and reorder the bag-of-words separately. The framework focuses on maximizing the occurrence of target words in the context of a given source sentence. Further, the permutation model used for reordering is still inferior to state-of-the-art reordering techniques. Hence, the lexical selection accuracy reported in this work is higher in comparison to the BLEU score. On the other hand, phrase-based translation produces a bag-of-phrases in the target language which are reordered using a distortion model. The framework focuses on maximizing the occurrence of target phrases in the context of source phrases and can potentially generate target hypotheses with both high lexical selection accuracy and BLEU score (weighted n-gram precision).

In the next section, we investigate the contribution of each dialog act to the overall improvement in translation quality. We analyze the performance in terms of lexical selection accuracy and BLEU score improvements per dialog act.

4.1. Analysis of results

Figure 3 shows the distribution of dialog acts in the 7 vocabulary dialog act tag set across the three corpora used in our experiments. *Statements* are the most frequent dialog acts followed by *question*, *other* and *acknowledgment*. Dialog acts such as *agreement, appreciation* and *abandoned* occur quite infrequently in the corpora.

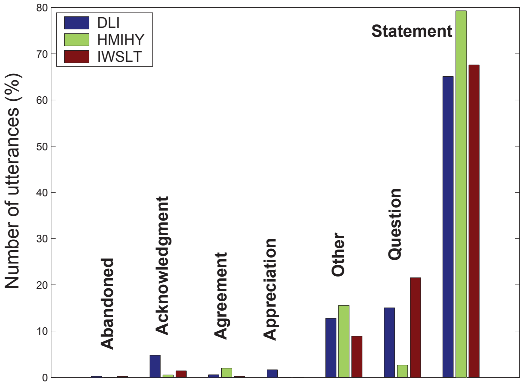

Fig. 3: Distribution of dialog acts in the test data of each corpus

In Table 4, we report the lexical selection accuracies and BLEU scores per dialog act for the BOW model and phrase-based translation model, respectively, on the Farsi-English corpus. The table compares the per DA performance of the two translation models with and without the use of dialog act information in the translation process. The results indicate that knowledge of discourse context such as *question* or *acknowledgment* is most beneficial to the translation process. Knowledge of detecting an utterance as a *statement* does not offer any significant improvement in the translation. This may be attributed to lack of systematic structural information (syntactic) or cue words that differentiate *statements* from other dialog acts. Deeper analysis using the 42 DA tag set indicates that dialog acts such as *yes-no questions*, *Wh-questions* and *open questions* contribute the most to the lexical selection accuracy and BLEU score improvement. Similar trends hold for the Chinese-English corpus. On the other hand, the improvements for the Japanese-English corpus is largely insignificant due to the high proportion of *statements* in the test corpus.

	BOW model		Phrase-based	
	Lexical accuracy		BLEU	
Dialog act	w/o DA	w/ DA	w/o DA	w/ DA
Statement	55.82	56.31	20.58	20.57
Question	59.85	62.14	24.12	26.36
Other	71.05	69.09	37.84	41.19
Acknowledgement	85.22	87.04	51.21	69.30
Appreciation	71.05	76.32	46.92	73.02
Agreement	56.00	66.67	18.46	50.00
Abandoned	75.00	75.00	58.41	58.41

Table 4: Lexical selection accuracy (%) and BLEU score (%) per DA tag for the BOW model and phrase-based translation scheme with and without use of dialog act tags for the DLI Farsi-English corpus

The analysis of the informativeness of dialog acts presented in this section has been performed only in terms of automatic evaluation metrics. As we have stressed before, the knowledge of dialog acts in translation may be much more beneficial in a cross-lingual human-computer or human-human interaction scenario that is not dependent on just word- (phrase-) level objective metrics. We plan to extend our analysis as part of our future work.

5. CONCLUSION AND FUTURE WORK

We have presented a new bag-of-words model for utilizing dialog act tags in spoken language translation. We contrasted the BOW model with a phrase-based interpolation scheme for integrating dialog acts from previous work. Our experiments indicate that exploiting DA tags with both the models provides promising improvements in terms of lexical selection accuracy and BLEU score. While integrating DA tags in the BOW model provides consistent improvement in lexical selection accuracy, it offers more pronounced improvements in BLEU score in the phrase table interpolation scheme.

While we have demonstrated that using dialog act tags can improve translation quality in terms of word based automatic evaluation metrics, the real benefits of such a scheme would be apparent through human evaluations. We are currently working on conducting subjective evaluations. The main objective of this paper was to demonstrate the utility of dialog acts in translation. Hence, the systems are not overly tuned or optimized to maximize the evaluation metrics. The results should be interpreted as a comparison between systems that do not have access to dialog acts with those that do have access to them. Furthermore, the experiments in this paper have been performed on reference transcripts. We plan to evaluate our framework on speech recognition output as well as word lattices as part of our future work.

6. REFERENCES

[1] N. Reithinger, R. Engel, M. Kipp, and M. Klesen, “Predicting dialogue acts for a speech-to-speech translation system,” in *Proc. of ICSLP*, Oct 1996, vol. 2, pp. 654–657.

[2] N. Reithinger and R. Engel, “Robust content extraction for translation and dialog processing,” in *Verbmobil: Foundations of Speech-to-Speech Translation*, W. Wahlster, Ed., pp. 430–439. Springer, 2000.

[3] H. Ney, F. J. Och, and S. Vogel, “Statistical translation of spoken dialogues in the verbmobil system,” in *Workshop on Multilingual Speech Communication*, Kyoto, 2000, pp. 69–74.

[4] A. Lavie, L. Levin, Y. Qu, A. Waibel, D. Gates, M. Gavalada, L. Mayfield, and M. Taboada, “Dialogue processing in a conversational speech translation system,” in *Proc. of ICSLP*, Oct 1996, vol. 1, pp. 554–557.

[5] E. Matusov, S. Kanthak, and H. Ney, “On the integration of speech recognition and statistical machine translation,” in *Proc. of Eurospeech*, 2005.

[6] L. Gu, Y. Gao, F. H. Liu, and M. Picheny, “Concept-based speech-to-speech translation using maximum entropy models for statistical natural concept generation,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 2, pp. 377–392, March 2006.

[7] P. D. Agüero, J. Adell, and A. Bonafonte, “Prosody generation for speech-to-speech translation,” in *Proceedings of ICASSP*, Toulouse, France, May 2006.

[8] V. K. Rangarajan Sridhar, S. Bangalore, and S. Narayanan, “Factored translation models for enriching spoken language translation with prosody,” in *Proceedings of Interspeech*, 2008.

[9] D. Jurafsky, R. Bates, N. Coccaro, R. Martin, M. Meteer, K. Ries, E. Shriberg, S. Stolcke, P. Taylor, and C. Van Ess-Dykema, “Switchboard discourse language modeling project report,” Technical report research note 30, Center for Speech and Language Processing, Johns Hopkins University, Baltimore, MD, 1998.

[10] E. Shriberg, R. Bates, A. Stolcke, P. Taylor, D. Jurafsky, K. Ries, N. Coccaro, R. Martin, M. Meteer, and C. Van Ess-Dykema, “Can prosody aid the automatic classification of dialog acts in conversational speech?,” *Language and Speech*, vol. 41, no. 3-4, pp. 439–487, 1998.

[11] V. K. Rangarajan Sridhar, S. Bangalore, and S. Narayanan, “Modeling the intonation of discourse segments for improved online dialog act tagging,” in *Proceedings of ICASSP*, Las Vegas, 2008.

[12] V. K. Rangarajan Sridhar, S. Bangalore, and S. Narayanan, “Enriching spoken language translation with dialog acts,” in *Proceedings of ACL*, 2008.

[13] S. Bangalore, P. Haffner, and S. Kanthak, “Statistical machine translation through global lexical selection and sentence reconstruction,” in *Proceedings of ACL*, 2007.

[14] P. Haffner, “Scaling large margin classifiers for spoken language understanding,” *Speech Communication*, vol. 48, no. iv, pp. 239–261, 2006.

[15] S. Narayanan et. al, “Speech recognition engineering issues in speech to speech translation system design for low resource languages and domains,” in *Proc. of ICASSP*, Toulose, France, May 2006.

[16] A. Gorin, G. Riccardi, and J. Wright, “How May I Help You?,” *Speech Communication*, vol. 23, pp. 113–127, 1997.

[17] M. Paul, “Overview of the IWSLT 2006 Evaluation Campaign,” in *Proc. of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2006, pp. 1–15.