Enriching spoken language translation with dialog acts

Vivek Kumar Rangarajan Sridhar
Shrikanth Narayanan
Speech Analysis and Interpretation Laboratory
University of Southern California
vrangara@usc.edu,shri@sipi.usc.edu

Srinivas Bangalore
AT&T Labs - Research
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 do not adequately utilize the rich contextual information such as conveyed through prosody and discourse function. In this paper, we explore the role of context characterized through *dialog acts* (DAs) in statistical translation. We demonstrate the integration of the dialog acts in a phrase-based statistical translation framework, employing 3 limited domain parallel corpora (Farsi-English, Japanese-English and Chinese-English). For all three language pairs, in addition to producing interpretable DA enriched target language translations, we also obtain improvements in terms of objective evaluation metrics such as lexical selection accuracy and BLEU score.

1 Introduction

Recent approaches to statistical speech translation have relied on improving translation quality with the use of phrase translation (Och and Ney, 2003; Koehn, 2004). The quality of phrase translation is typically measured using n-gram precision based metrics such as BLEU (Papineni et al., 2002) and NIST scores. However, in many dialog based speech translation scenarios, vital information beyond what is robustly captured by words and phrases is carried by the communicative act (e.g., *question*, *acknowledgement*, etc.) representing the function of the utterance. Our approach for incorporating dialog act tags in speech translation is motivated by the fact that it is important to capture and convey not only *what* is being communicated (the words) but *how* something is being communicated (the context). Augmenting current statistical translation frameworks with *dialog acts* can potentially improve translation quality and facilitate successful cross-lingual interactions in terms of improved information transfer.

Dialog act tags have been previously used in the VERBMOBIL statistical speech-to-speech translation system (Reithinger et al., 1996). In that work, the predicted DA tags were mainly used to improve speech recognition, semantic evaluation, and information extraction modules. Discourse information in the form of speech acts has also been used in interlingua translation systems (Mayfield et al., 1995) to map input text to semantic concepts, which are then translated to target text.

In contrast with previous work, in this paper we demonstrate how dialog act tags can be directly exploited in phrase based statistical speech translation systems (Koehn, 2004). The framework presented in this paper is particularly suited for human-human and human-computer interactions in a dialog setting, where information loss due to erroneous content may be compensated to some extent through the correct transfer of the appropriate dialog act. The dialog acts can also be potentially used for imparting correct utterance level intonation during speech synthesis in the target language. Figure 1 shows an example where the detection and transfer of dialog act information is beneficial in resolving ambiguous intention associated with the translation output.

Source: آیا این مسکنه
Ref: is this a painkiller
Hyp: this is a painkiller
Enriched Hyp: this is a painkiller (Yes-No-Question)

Figure 1: Example of speech translation output enriched with dialog act

The remainder of this paper is organized as follows: Section 2 describes the dialog act tagger used in this work, Section 3 formulates the problem, Section 4 describes the parallel corpora used in our experiments, Section 5 summarizes our experimental results and Section 6 concludes the paper with a brief discussion and outline for future work.

2 Dialog act tagger

In this work, we use a dialog act tagger trained on the Switchboard DAMSL corpus (Jurafsky et al.,

Proceedings of ACL-08: HLT, Short Papers (Companion Volume), pages 225–228,
Columbus, Ohio, USA, June 2008.

1998) using a maximum entropy (maxent) model. The Switchboard-DAMSL (SWBD-DAMSL) corpus consists of 1155 dialogs and 218,898 utterances from the Switchboard corpus of telephone conversations, tagged with discourse labels from a shallow discourse tagset. The original tagset of 375 unique tags was clustered to obtain 42 dialog tags as in (Jurafsky et al., 1998). In addition, we also grouped the 42 tags into 7 disjoint classes, based on the frequency of the classes and grouped the remaining classes into an "Other" category constituting less than 3% of the entire data. The simplified tagset consisted of the following classes: *statement, acknowledgment, abandoned, agreement, question, appreciation, other*.

We use a maximum entropy sequence tagging model for the automatic DA tagging. Given a sequence of utterances $U = u_1, u_2, \cdots, u_n$ and a dialog act vocabulary ($d_i \in \mathcal{D}, |\mathcal{D}| = K$), we need to assign the best dialog act sequence $D^* = d_1, d_2, \cdots, d_n$. The classifier is used to assign to each utterance a dialog act label conditioned on a vector of local contextual feature vectors comprising the lexical, syntactic and acoustic information. We used the machine learning toolkit LLAMA (Haffner, 2006) to estimate the conditional distribution using maxent. The performance of the maxent dialog act tagger on a test set comprising 29K utterances of SWBD-DAMSL is shown in Table 1.

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.

3 Enriched translation using DAs

If S_s, T_s and S_t, T_t are the speech signals and equivalent textual transcription in the source and target language, and L_s the enriched representation for the source speech, we formalize our proposed enriched S2S translation in the following manner:

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

$$P(S_t|S_s) = \sum_{T_t,T_s,L_s} P(S_t, T_t, T_s, L_s|S_s) \quad (2)$$

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

where Eq.(3) is obtained through conditional independence assumptions. Even though the recognition and translation can be performed jointly (Matusov et al., 2005), typical S2S translation frameworks compartmentalize the ASR, MT and TTS, with each component maximized for performance individually.

$$\begin{aligned}\max_{S_t} P(S_t|S_s) \approx & \max_{S_t} P(S_t|T_t^*, L_s^*) \\ & \times \max_{T_t} P(T_t|T_s^*, L_s^*) \\ & \times \max_{L_s} P(L_s|T_s^*, S_s) \times \max_{T_s} P(T_s|S_s)\end{aligned} \quad (4)$$

where T_s^*, T_t^* and S_t^* are the arguments maximizing each of the individual components in the translation engine. 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 (Gu et al., 2006), prosody (Agüero et al., 2006), or, as in this work, dialog act tags.

3.1 Phrase-based translation with dialog acts

One of the currently popular and predominant schemes for statistical translation is the phrase-based approach (Koehn, 2004). Typical phrase-based SMT approaches obtain word-level alignments from a bilingual corpus using tools such as GIZA++ (Och and Ney, 2003) and extract phrase translation pairs from the bilingual word alignment using heuristics. Suppose, the SMT had access to source language dialog acts (L_s), the translation problem may be reformulated as,

$$\begin{aligned}T_t^* &= \arg\max_{T_t} P(T_t|T_s, L_s) \\ &= \arg\max_{T_t} P(T_s|T_t, L_s).P(T_t|L_s)\end{aligned} \quad (5)$$

The first term in Eq.(5) corresponds to a dialog act specific MT model and the second term to a dialog act specific language model. Given sufficient amount of training data such a system can possibly generate hypotheses that are more accurate than the scheme without the use of dialog acts. However, for small scale and limited domain applications, Eq.(5) leads to an implicit partitioning of the data corpus

	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.

and might generate inferioir translations in terms of lexical selection accuracy or BLEU score.

A natural step to overcome the sparsity issue is to employ an appropriate back-off mechanism that would exploit the phrase translation pairs derived from the complete data. A typical phrase translation table consists of 5 phrase translation scores for each pair of phrases, source-to-target phrase translation probability (λ_1), target-to-source phrase translation probability (λ_2), source-to-target lexical weight (λ_3), target-to-word lexical weight (λ_4) and phrase penalty (λ_5= 2.718). The lexical weights are the product of word translation probabilities obtained from the word alignments. 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 phrase table of the DA-specific translation model. The appended entries are weighted by a factor α.

$$(T_s \rightarrow T_t)_{L_s^*} = (T_s \rightarrow T_t)_{L_s} \cup \{\alpha.(T_s \rightarrow T_t) \; s.t. \; (T_s \rightarrow T_t) \notin (T_s \rightarrow T_t)_{L_s}\} \quad (6)$$

where $(T_s \rightarrow T_t)$ is a short-hand[1] notation for a phrase translation table. $(T_s \rightarrow T_t)_{L_s}$ is the DA-specific phrase translation table, $(T_s \rightarrow T_t)$ is the phrase translation table constructed from entire 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. Such a scheme ensures that phrase translation pairs belonging to a specific DA model are weighted higher and also ensures better coverage than a partitioned data set.

4 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 human-mediated doctor-patient mediated interactions in which an English speaking doctor interacts with a Persian speaking patient (Narayanan et al., 2006). We used a subset of this corpus consisting of 9315 parallel sentences.

The Japanese-English parallel corpus is a part of the "How May I Help You" (HMIHY) (Gorin et al., 1997) corpus of operator-customer conversations related to telephone services. The corpus consists of 12239 parallel sentences. The conversations are spontaneous even though the domain is limited. The Chinese-English corpus corresponds to the IWSLT06 training and 2005 development set comprising 46K and 506 sentences respectively (Paul, 2006).

5 Experiments and Results

In all our experiments we assume that the same dialog act is shared by a parallel sentence pair. Thus, even though the dialog act prediction is performed for English, we use the predicted dialog act as the dialog act for the source language sentence. We used the Moses[2] toolkit for statistical phrase-based translation. The language models were trigram models created only from the training portion of each corpus. Due to the relatively small size of the corpora used in the experiments, we could not devote a separate development set for tuning the parameters of the phrase-based translation scheme. Hence, the experiments are strictly performed on the training and test sets reported in Table 2[3].

The lexical selection accuracy and BLEU scores for the three parallel corpora is 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

[1] $(T_s \rightarrow T_t)$ 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).

[2] http://www.statmt.org/moses

[3] A very small subset of the data was reserved for optimizing the interpolation factor (α) described in Section 3.1

Language pair	F-score (%)			BLEU (%)		
	w/o DA tags	w/ DA tags		w/o DA tags	w/ DA tags	
		7tags	42tags		7tags	42tags
Farsi-English	56.46	57.32	57.74	22.90	23.50	23.75
Japanese-English	79.05	79.40	79.51	54.15	54.21	54.32
Chinese-English	65.85	67.24	67.49	48.59	52.12	53.04

Table 3: F-measure and BLEU scores with and without use of dialog act tags.

the set of words in the translation output. 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. The improvements for Farsi-English and Chinese-English corpora are more pronounced than the improvements in Japanese-English corpus. This is due to the skewed distribution of dialog acts in the Japanese-English corpus; 80% of the test data are *statements* while *other* and *questions* category make up 16% and 3.5% of the data respectively. The important observation here is that, appending DA tags in the form described in this work, can improve translation performance even in terms of conventional objective evaluation metrics. However, the performance gain measured in terms of objective metrics that are designed to reflect only the orthographic accuracy during translation is not a complete evaluation of the translation quality of the proposed framework. We are currently planning of adding human evaluation to bring to fore the usefulness of such rich annotations in interpreting and supplementing typically noisy translations.

6 Discussion and Future Work

It is important to note that the dialog act tags used in our translation system are predictions from the maxent based DA tagger described in Section 2. We do not have access to the reference tags; thus, some amount of error is to be expected in the DA tagging. Despite the lack of reference DA tags, we are still able to achieve modest improvements in the translation quality. Improving the current DA tagger and developing suitable adaptation techniques are part of future work.

While we have demonstrated here 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 attested through further human evaluations. We are currently working on conducting subjective evaluations.

References

P. D. Agüero, J. Adell, and A. Bonafonte. 2006. Prosody generation for speech-to-speech translation. In *Proc. of ICASSP*, Toulouse, France, May.

A. Gorin, G. Riccardi, and J. Wright. 1997. How May I Help You? *Speech Communication*, 23:113–127.

L. Gu, Y. Gao, F. H. Liu, and M. Picheny. 2006. Concept-based speech-to-speech translation using maximum entropy models for statistical natural concept generation. *IEEE Transactions on Audio, Speech and Language Processing*, 14(2):377–392, March.

P. Haffner. 2006. Scaling large margin classifiers for spoken language understanding. *Speech Communication*, 48(iv):239–261.

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

P. Koehn. 2004. Pharaoh: A beam search decoder for phrasebased statistical machine translation models. In *Proc. of AMTA-04*, pages 115–124.

E. Matusov, S. Kanthak, and H. Ney. 2005. On the integration of speech recognition and statistical machine translation. In *Proc. of Eurospeech*.

L. Mayfield, M. Gavalda, W. Ward, and A. Waibel. 1995. Concept-based speech translation. In *Proc. of ICASSP*, volume 1, pages 97–100, May.

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

F. J. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

K. Papineni, S. Roukos, T. Ward, and W. J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. Technical report, IBM T.J. Watson Research Center.

M. Paul. 2006. Overview of the IWSLT 2006 Evaluation Campaign. In *Proc. of the IWSLT*, pages 1–15, Kyoto, Japan.

N. Reithinger, R. Engel, M. Kipp, and M. Klesen. 1996. Predicting dialogue acts for a speech-to-speech translation system. In *Proc. of ICSLP*, volume 2, pages 654–657, Oct.