Incremental Topic-Based Translation Model Adaptation for Conversational Spoken Language Translation

Sanjika Hewavitharana, Dennis N. Mehay, Sankaranarayanan Ananthakrishnan and Prem Natarajan

Speech, Language and Multimedia Business Unit
Raytheon BBN Technologies
Cambridge, MA 02138, USA
{shewavit,dmehay,sanantha,pnataraj}@bbn.com

Abstract

We describe a translation model adaptation approach for conversational spoken language translation (CSLT), which encourages the use of contextually appropriate translation options from relevant training conversations. Our approach employs a monolingual LDA topic model to derive a similarity measure between the test conversation and the set of training conversations, which is used to bias translation choices towards the current context. A significant novelty of our adaptation technique is its *incremental* nature; we continuously update the topic distribution on the evolving test conversation as new utterances become available. Thus, our approach is well-suited to the causal constraint of spoken conversations. On an English-to-Iraqi CSLT task, the proposed approach gives significant improvements over a baseline system as measured by BLEU, TER, and NIST. Interestingly, the incremental approach outperforms a non-incremental oracle that has up-front knowledge of the whole conversation.

1 Introduction

Conversational spoken language translation (CSLT) systems facilitate communication between subjects who do not speak the same language. Current systems are typically used to achieve a specific task (e.g. vehicle checkpoint search, medical diagnosis, etc.). These task-driven conversations typically revolve around a set of central topics, which may not be evident at the beginning of the interaction. As the conversation progresses, however, the gradual accumulation of contextual information can be used to infer the topic(s) of discussion, and to deploy contextually appropriate translation phrase pairs. For example, the word '*drugs*' will predominantly translate into Spanish as '*medicamentos*' (medicines) in a medical scenario, whereas the translation '*drogas*' (illegal drugs) will predominate in a law enforcement scenario. Most CSLT systems do not take high-level global context into account, and instead translate each utterance in isolation. This often results in contextually inappropriate translations, and is particularly problematic in conversational speech, which usually exhibits short, spontaneous, and often ambiguous utterances.

In this paper, we describe a novel topic-based adaptation technique for phrase-based statistical machine translation (SMT) of spoken conversations. We begin by building a monolingual latent Dirichlet allocation (LDA) topic model on the training conversations (each conversation corresponds to a "document" in the LDA paradigm). At run-time, this model is used to infer a topic distribution over the evolving test conversation up to and including the current utterance. Translation phrase pairs that originate in training conversations whose topic distribution is similar to that of the current conversation are given preference through a single similarity feature, which augments the standard phrase-based SMT log-linear model. The topic distribution for the test conversation is updated incrementally for each new utterance as the available history grows. With this approach, we demonstrate significant improvements over a baseline phrase-based SMT system as measured by BLEU, TER and NIST scores on an English-to-Iraqi CSLT task.

Disclaimer: This paper is based upon work supported by the DARPA BOLT program. The views expressed here are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

Distribution Statement A (Approved for Public Release, Distribution Unlimited)

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 697–701,
Sofia, Bulgaria, August 4-9 2013.

2 Relation to Prior Work

Domain adaptation to improve SMT performance has attracted considerable attention in recent years (Foster and Kuhn, 2007; Finch and Sumita, 2008; Matsoukas et al., 2009). The general theme is to divide the training data into partitions representing different domains, and to prefer translation options for a test sentence from training domains that most resemble the current document context. Weaknesses of this approach include (a) assuming the existence of discrete, non-overlapping domains; and (b) the unreliability of models generated by segments with little training data.

To avoid the need for hard decisions about domain membership, some have used topic modeling to improve SMT performance, e.g., using latent semantic analysis (Tam et al., 2007) or 'biTAM' (Zhao and Xing, 2006). In contrast to our source language approach, these authors use both source and target information.

Perhaps most relevant are the approaches of Gong et al. (2010) and Eidelman et al. (2012), who both describe adaptation techniques where monolingual LDA topic models are used to obtain a topic distribution over the training data, followed by dynamic adaptation of the phrase table based on the inferred topic of the test document. While our proposed approach also employs monolingual LDA topic models, it deviates from the above methods in the following important ways. First, the existing approaches are geared towards batch-mode text translation, and assume that the full document context of a test sentence is always available. This assumption is incompatible with translation of spoken conversations, which are inherently causal. Our proposed approach infers topic distributions *incrementally* as the conversation progresses. Thus, it is not only consistent with the causal requirement, but is also capable of tracking topical changes during the course of a conversation.

Second, we do not directly augment the translation table with the inferred topic distribution. Rather, we compute a similarity between the current conversation history and each of the training conversations, and use this measure to dynamically score the relevance of candidate translation phrase pairs during decoding.

3 Corpus Data and Baseline SMT

We use the DARPA TransTac English-Iraqi parallel two-way spoken dialogue collection to train both translation and LDA topic models. This data set contains a variety of scenarios, including medical diagnosis; force protection (e.g. checkpoint, reconnaissance, patrol); aid, maintenance and infrastructure, etc.; each transcribed from spoken bilingual conversations and manually translated. The SMT parallel training corpus contains approximately 773K sentence pairs (7.3M English words). We used this corpus to extract translation phrase pairs from bidirectional IBM Model 4 word alignment (Och and Ney, 2003) based on the heuristic approach of (Koehn et al., 2003). A 4-gram target LM was trained on all Iraqi Arabic transcriptions. Our phrase-based decoder is similar to Moses (Koehn et al., 2007) and uses the phrase pairs and target LM to perform beam search stack decoding based on a standard log-linear model, the parameters of which were tuned with MERT (Och, 2003) on a held-out development set (3,534 sentence pairs, 45K words) using BLEU as the tuning metric. Finally, we evaluated translation performance on a separate, unseen test set (3,138 sentence pairs, 38K words).

Of the 773K training sentence pairs, about 100K (corresponding to 1,600 conversations) are marked with conversation boundaries. We use the English side of these conversations for training LDA topic models. All other sentence pairs are assigned to a "background conversation", which signals the absence of the topic similarity feature for phrase pairs derived from these instances. All of the development and test set data were marked with conversation boundaries. The training, development and test sets were partitioned at the conversation level, so that we could model a topic distribution for entire conversations, both during training and during tuning and testing.

4 Incremental Topic-Based Adaptation

Our approach is based on the premise that biasing the translation model to favor phrase pairs originating in training conversations that are contextually similar to the current conversation will lead to better translation quality. The topic distribution is incrementally updated as the conversation history grows, and we recompute the topic similarity between the current conversation and the training conversations for each new source utterance.

4.1 Topic modeling with LDA

We use latent Dirichlet allocation, or LDA, (Blei et al., 2003) to obtain a topic distribution over conversations. For each conversation d_i in the training collection (1,600 conversations), LDA infers a topic distribution $\theta_{d_i} = p(z_k|d_i)$ for all latent topics $z_k = \{1, ..., K\}$, where K is the number of topics. In this work, we experiment with values of $K \in \{20, 30, 40\}$. The full conversation history is available for training the topic models and estimating topic distributions in the training set.

At run-time, however, we construct the conversation history for the tuning and test sets incrementally, one utterance at a time, mirroring a real-world scenario where our knowledge is limited to the utterances that have been spoken up to that point in time. Thus, each development/test utterance is associated with a different conversation history d^*, for which we infer a topic distribution $\theta_{d^*} = p(z_k|d^*)$ using the trained LDA model. We use Mallet (McCallum, 2002) for training topic models and inferring topic distributions.

4.2 Topic Similarity Computation

For each test utterance, we are able to infer the topic distribution θ_{d^*} based on the accumulated history of the current conversation. We use this to compute a measure of similarity between the evolving test conversation and each of the training conversations, for which we already have topic distributions θ_{d_i}. Because θ_{d_i} and θ_{d^*} are probability distributions, we use the Jensen-Shannon divergence (JSD) to evaluate their similarity (Manning and Schütze, 1999). The JSD is a smoothed and symmetric version of Kullback-Leibler divergence, which is typically used to compare two probability distributions. We define the similarity score as $sim(\theta_{d_i}, \theta_{d^*}) = 1 - JSD(\theta_{d_i}||\theta_{d^*})$.[1] Thus, we obtain a vector of similarity scores indexed by the training conversations.

4.3 Integration with the Decoder

We provide the SMT decoder with the similarity vector for each test utterance. Additionally, the SMT phrase table tracks, for each phrase pair, the set of parent training conversations (including the "background conversation") from which that phrase pair originated. Using this information, the decoder evaluates, for each candidate phrase pair $X \rightarrow Y$ added to the search graph, its topic similarity score as follows:

$$F_{X \rightarrow Y} = \max_{i \in Par(X \rightarrow Y)} sim(\theta_{d_i}, \theta_{d^*}) \quad (1)$$

where $Par(X \rightarrow Y)$ is the set of training conversations from which the candidate phrase pair originated. Phrase pairs from the "background conversation" only are assigned a similarity score $F_{X \rightarrow Y} = 0.00$. In this way we distill the inferred topic distributions down to a single feature for each candidate phrase pair. We add this feature to the log-linear translation model with its own weight, which is tuned with MERT. The intuition behind this feature is that the lower bound of suitability of a candidate phrase pair should be directly proportional to the similarity between its most relevant conversational provenance and the current context. Phrase pairs which only occur in the background conversation are not directly penalized, but contribute nothing to the topic similarity score.

[1] JSD($\theta_{d_i}||\theta_{d^*}$) $\in [0, 1]$ when defined using $\log_2$.

	REFERENCE TRANSCRIPTIONS		
SYSTEM	BLEU↑	TER↓	NIST↑
Baseline	19.32	58.66	6.22
incr20	19.39	58.44	6.26*
incr30	19.36	58.32*	6.26
incr40	**19.68***	**58.19***	**6.28***
conv20	19.60*	58.36*	6.27*
conv30	19.48	58.38*	6.27*
conv40	19.50	58.33*	6.28*
	ASR TRANSCRIPTIONS		
SYSTEM	BLEU↑	TER↓	NIST↑
Baseline	16.92	62.57	5.75
incr20	16.99	62.28*	5.77
incr30	16.96	62.33*	5.78
incr40	**17.31***	**61.97***	**5.83***
conv20	17.29*	62.28*	5.81*
conv30	17.12	62.19*	5.80*
conv40	17.00	62.14*	5.79*

Table 1: Stemmed results on 3,138-utterance test set. Asterisked results are significantly better than the baseline ($p \leq 0.05$) using 1,000 iterations of paired bootstrap re-sampling (Koehn, 2004). (**Key:** incr$N$ = incremental LDA with N topics; conv$N$ = non-incremental, whole-conversation LDA with N topics.)

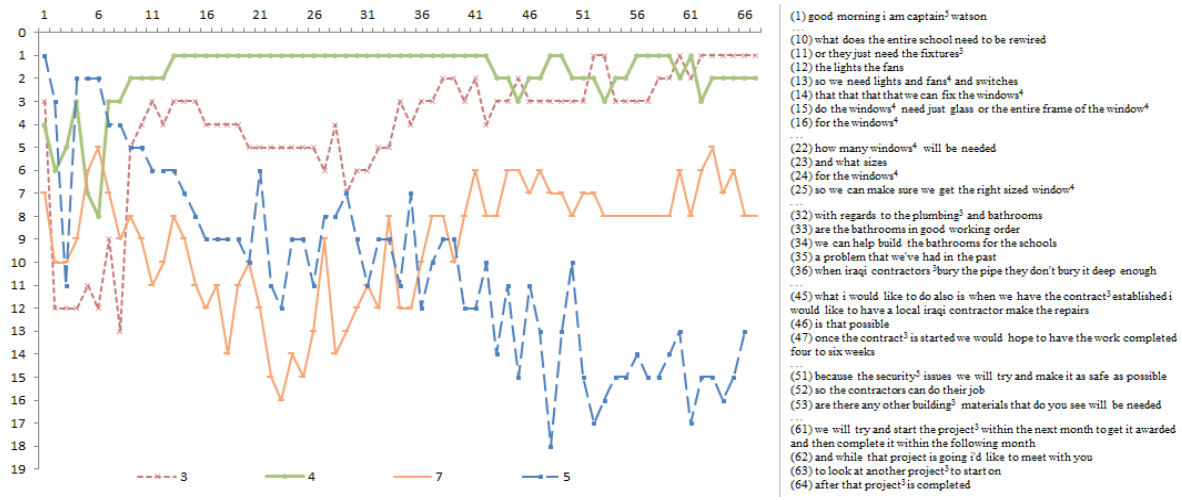

Figure 1: Rank trajectories of 4 LDA inferred topics, with incremental topic inference. The x-axis indicates the utterance number. The y-axis indicates a topic's rank at each utterance.

5 Experimental Setup and Results

The baseline English-to-Iraqi phrase-based SMT system was built as described in Section 3. This system translated each utterance independently, ignoring higher-level conversational context.

For the topic-adapted system, we compared translation performance with a varying number of LDA topics. In intuitive agreement with the approximate number of scenario types known to be covered by our data set, a range of 20-40 topics yielded the best results. We compared the proposed incremental topic tracking approach to a non-causal oracle approach that had up-front access to the entire source conversations at run-time.

In all cases, we compared translation performance on both clean-text and automatic speech recognition (ASR) transcriptions of the source utterances. ASR transcriptions were generated using a high-performance two-pass HMM-based system, which delivered a word error rate (WER) of 10.6% on the test set utterances.

Table 1 summarizes test set performance in BLEU (Papineni et al., 2001), NIST (Doddington, 2002) and TER (Snover et al., 2006). Given the morphological complexity of Iraqi Arabic, computing string-based metrics on raw output can be misleadingly low and does not always reflect whether the core message was conveyed. Since the primary goal of CSLT is information transfer, we present automatic results that are computed after stemming with an Iraqi Arabic stemmer.

We note that in all settings (incremental and non-causal oracle) our adaptation approach matches or significantly outperforms the baseline across multiple evaluation metrics. In particular, the incremental LDA system with 40 topics is the top-scoring system in both clean-text and ASR settings. In the ASR setting, which simulates a real-world deployment scenario, this system achieves improvements of 0.39 (BLEU), -0.6 (TER) and 0.08 (NIST).

6 Discussion and Future Directions

We have presented a novel, incremental topic-based translation model adaptation approach that obeys the causality constraint imposed by spoken conversations. This approach yields statistically significant gains in standard MT metric scores.

We have also demonstrated that incremental adaptation on an evolving conversation performs better than oracle adaptation based on the complete conversation history. Although this may seem counter-intuitive, Figure 1 gives clues as to why this happens. This figure illustrates the rank trajectory of four LDA topics as the incremental conversation grows. The accompanying text shows excerpts from the conversation. We indicate (in superscript) the topic identity of most relevant words in an utterance that are associated with that topic. At the first utterance, the top-ranked topic is "5", due to the occurrence of "captain" in the greeting. As the conversation evolves, we note that this topic become less prominent. The conversation shifts to a discussion on "windows", raising the prominence of topic "4". Finally, topic "3" becomes prominent due to the presence of the

words “project” and “contract”. Thus, the incremental approach is able to track the topic trajectories in the conversation, and is able to select more relevant phrase pairs than oracle LDA, which estimates one topic distribution for the entire conversation.

In this work we have used only the source language utterance in inferring the topic distribution. In a two-way CLST system, we also have access to SMT-generated back-translations in the Iraqi-English direction. As a next step, we plan to use SMT-generated English translation of Iraqi utterances to improve topic estimation.

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, March.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, HLT ’02, pages 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Vladimir Eidelman, Jordan Boyd-Graber, and Philip Resnik. 2012. Topic models for dynamic translation model adaptation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL ’12, pages 115–119, Stroudsburg, PA, USA. Association for Computational Linguistics.

Andrew Finch and Eiichiro Sumita. 2008. Dynamic model interpolation for statistical machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, StatMT ’08, pages 208–215, Stroudsburg, PA, USA. Association for Computational Linguistics.

George Foster and Roland Kuhn. 2007. Mixture-model adaptation for SMT. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT ’07, pages 128–135, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zhengxian Gong, Yu Zhang, and Guodong Zhou. 2010. Statistical machine translation based on LDA. In *Universal Communication Symposium (IUCS), 2010 4th International*, pages 286–290.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *NAACL ’03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54, Morristown, NJ, USA. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL ’07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *EMNLP*, pages 388–395, Barcelona, Spain, July.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA, USA.

Spyros Matsoukas, Antti-Veikko I. Rosti, and Bing Zhang. 2009. Discriminative corpus weight estimation for machine translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP ’09, pages 708–717, Stroudsburg, PA, USA. Association for Computational Linguistics.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, March.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *ACL ’03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 160–167, Morristown, NJ, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: A method for automatic evaluation of machine translation. In *ACL ’02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Morristown, NJ, USA. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings AMTA*, pages 223–231, August.

Yik-Cheung Tam, Ian Lane, and Tanja Schultz. 2007. Bilingual LSA-based adaptation for statistical machine translation. *Machine Translation*, 21(4):187–207, December.

Bing Zhao and Eric P. Xing. 2006. BiTAM: Bilingual topic admixture models for word alignment. In *In Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL ‘06)*.