Anticipatory Translation Model Adaptation for Bilingual Conversations

Sanjika Hewavitharana, Dennis Mehay, Sankaranarayanan Ananthakrishnan,
Rohit Kumar, John Makhoul

Speech, Language and Multimedia Business Unit
Raytheon BBN Technologies
Cambridge, MA 02138, U.S.A
{shewavit,dmehay,sanantha,rkumar,makhoul}@bbn.com

Abstract

Conversational spoken language translation (CSLT) systems facilitate bilingual conversations in which the two participants speak different languages. Bilingual conversations provide additional contextual information that can be used to improve the underlying machine translation system. In this paper, we describe a novel translation model adaptation method that *anticipates* a participant's response in the target language, based on his counterpart's prior turn in the source language. Our proposed strategy uses the source language utterance to perform cross-language retrieval on a large corpus of bilingual conversations in order to obtain a set of potentially relevant target responses. The responses retrieved are used to bias translation choices towards anticipated responses. On an Iraqi-to-English CSLT task, our method achieves a significant improvement over the baseline system in terms of BLEU, TER and METEOR metrics.

1. Introduction

State of the art conversational spoken language translation (CSLT) systems enable useful, functional communication between two subjects who do not speak the same language. In a typical CSLT pipeline, source language speech is transcribed using automatic speech recognition (ASR), piped to text-to-text statistical machine translation (SMT), followed by text-to-speech (TTS) synthesis in the target language. Two sets of these components are used; one in the source-to-target direction and another in the target-to-source direction. The two directions are typically processed independently, where successive turns in the source and target languages are processed in complete isolation. This decoupling sometimes leads to contextually inappropriate translations.

Fortunately, bilingual conversations offer a wealth of contextual information that can be exploited to improve translation performance. Contextual cues can be used to adapt the translation model and improve its relevance to the current state of the dialogue. Typically, the adaptation is done monolingually, using only the utterances of one speaker. In this paper, we describe a novel translation model adaptation technique for bilingual conversations that *anticipates* a participant's response in the target language based on his *counterpart's* prior turn in the source language. Depending on the nature of the bilingual conversation, adaptation can be profitably performed in either language. We evaluate the proposed approach on Iraqi-English bilingual conversations drawn from the DARPA TransTac/BOLT spoken dialogue collection.

Our approach is motivated by the observation that in most domains, the primary goal of bilingual conversations is to exchange information across the language barrier. To that end, the most useful translation is often the one that most effectively conveys the content of a speaker's *response* to the content of the *counterpart's preceding utterances*. Table 1 illustrates this with an excerpt from an Iraqi-English bilingual conversation at a vehicle checkpoint from the DARPA TransTac/BOLT corpus. The first column corresponds to the English speaker's turn; the second column is the Iraqi speaker's following turn, or response (in Buckwalter transliteration); the third column provides an English gloss of the Iraqi speaker's response. As in most cooperative conversations, the Iraqi responses are all relevant to the preceding English turn, and, in many cases, largely predictable from the preceding English turn in the first column.

Following these observations, we perform turn-level translation model adaptation that prefers phrasal translation rules that originate from *responses* that immediately follow *counterpart utterances* that are similar to those of the *current conversational counterpart*. This approach produces a measurable improvement over a phrase-based SMT baseline system in terms of BLEU, TER and METEOR metrics on an Iraqi-to-English translation task.

2. Anticipatory Translation Model Adaptation

Our adaptation scheme attempts to model the effect of the preceding target language turn on the translation of the current source language utterance. The intuition is that biasing

This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA). The views, opinions, and/or findings contained in this article are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.
Approved for Public Release, Distribution Unlimited

N	English Turn	Iraqi Response	English Gloss of Iraqi Response
1.	*turn off your engine and get out of the car*	*tfDlwA*	*here you are*
2.	*give me your i_d*	*bTAqty wjwAzy*	*my i_d card and my passport*
3.	*where you coming from*	*mn swryA mn dyr Alzwr*	*from syria from dair al-zour*
4.	*and where you going*	*rAyH llrmAdy*	*i'm going to ramadi*
5.	*what's in your truck*	*Iny bqAl wdJjyb xs JbyEh hnAk*	*i'm a grocer and i'm bringing lettuce*

Table 1: Excerpt from an Iraqi-English bilingual conversation in the DARPA TransTac/BOLT collection.

the translation model to favor phrase pairs originating from training utterances that have similar preceding target language turn will produce translations more appropriate to the current conversation. Such a model can be learned in a data-driven fashion from a large training corpus of bilingual conversations, organized in the form of starting target language turns and ensuing source language responses. The DARPA TransTac/BOLT spoken dialogue corpus is organized as a collection of bilingual conversations, thus making it relatively simple to build an "anticipatory parallel corpus" (APC) of target language turn and source language response pairs for training the translation model (see Section 3). The APC is a pseudo-parallel corpus with prior target turns mapping to the immediately following source language responses, similar to the first two columns of Table 1. In the following description, we assume, without loss of generality, that we are performing cross-lingual translation model adaptation for translating the current Iraqi turn into English based on the preceding English turn. Figure 1 illustrates the adaptation process.

2.1. Cross-Lingual Retrieval

When decoding the current Iraqi utterance in the context of a bilingual conversation, we seek to predict what an appropriate response to the preceding English turn might look like.[1] To find support for this prediction, we use the preceding English turn as a query to perform cross-lingual retrieval on the APC constructed from the training conversations. The goal of this step is to obtain the most relevant Iraqi responses to the preceding English turn. Each training utterance pair in the APC is assigned a unique utterance ID, which we later use in the online adaptation of the translation model (Section 2.2).

Because the APC is not a true parallel corpus in the sense that the Iraqi responses are not direct translations of the preceding English turns, learning a true cross-lingual retrieval model from this data would be difficult. Instead, we employed the simpler approach of first performing *monolingual* retrieval of the English turns most similar to the query turn, and then reading off the corresponding Iraqi responses from the APC. To facilitate this, we represent all APC English turns in a trigram term-indicator vector space with appropriate pre-processing (e.g. stop-word removal), and we index each training utterance separately. During retrieval, we map the preceding English turn to the same vector space, and select APC English turns that have the largest cosine similarity to the query. We then read off the corresponding Iraqi response turns from the APC. This produces a *bias corpus* of Iraqi responses that might be relevant to the preceding English turn, which we limit to a small number between 50 and 500 in our experiments.

Table 2 illustrates anticipatory cross-lingual retrieval with an example. The first row corresponds to the query English turn. The first column of the second row lists the five top-ranking Iraqi responses retrieved from the APC using the above mechanism. The second column of the second row provides an English gloss for the retrieved Iraqi responses. The final row shows the actual Iraqi response to the query English turn, and its English gloss. In this example, the retrieved Iraqi responses are well-matched to the actual response. Thus, a translation model biased towards the phrases extracted from the retrieved responses is likely to produce better translations.

Q.	**how are you doing today**	
1.	*wAllh AlHmd llh zyn*	*well fine thank god*
2.	*SbAH Alnwr JhlAF wshlAF*	*good morning hello and welcome*
3.	*JhlAF byk kyf AlHAl*	*hello to you how are you*
4.	*SbAH Alxyr JhlAF wshlAF AlHmd llh zyn*	*good morning and welcome thank god i'm well*
5.	*Iny zyn JHsn mn Endh*	*i'm fine better than him*
R.	**AlHmd llh zyn**	**good thank god**

Table 2: Iraqi response retrieval for a sample English query turn.

2.2. Translation Model Adaptation

From the cross-lingual retrieval on each previous English turn, we obtain for each Iraqi turn I, a set of anticipated Iraqi responses, corresponding Iraqi utterance IDs and a set of similarity scores (cosine similarity between the query English turn and APC English turns) $\mathbf{R}$. We use these scores directly as relevance scores for the anticipatory Iraqi responses. At run time, an updated relevance vector is passed on to the

[1] In an interactive CSLT system, all utterances by speakers of both languages are transcribed by speech recognition (ASR), though we also run oracle experiments with the ground truth (reference) transcription.

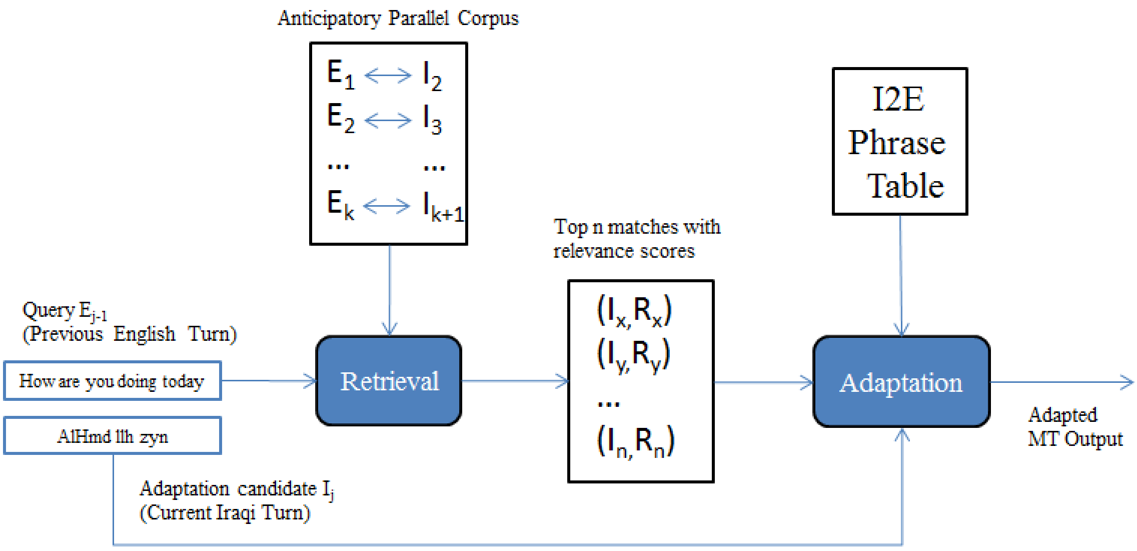

Figure 1: Anticipatory translation model adaptation process.

SMT decoder for each new test utterance.

The SMT phrase table tracks, for each phrase pair, the set of training utterances from which that phrase pair originated. Only part of the training corpus has marked conversation boundaries. Phrase translation rules derived from sentence pairs that do not originate in bilingual conversations are assigned a default utterance ID. For each candidate phrase pair $\overline{I} \rightarrow \overline{E}$ added to the search graph, the SMT decoder computes the relevance score as the maximum of all relevance scores corresponding to the current turn. i.e.

$$F_{\overline{I}\rightarrow\overline{E}} = \max_{j\in Par(\overline{I}\rightarrow\overline{E})} \mathbf{R}_j \qquad (1)$$

where $Par(\overline{I} \rightarrow \overline{E})$ is the set of training utterances from which the candidate phrase pair originated. Phrase pairs with the default utterance ID are assigned a default relevance score of 0.0. (in effect, they are decoded with the baseline features only). The relevance score is added as a feature to the log-linear translation model with its own weight, which is tuned with the rest of the parameters. The effect of this feature is to bias the decoder in favor of phrase pairs that originate in relevant responses.

3. Baseline SMT System

We use the DARPA TransTac/BOLT Iraqi-English parallel two-way spoken dialogue collection to train the translation models. Each conversation represents an interaction between an English interviewer and an Iraqi respondent, based on a scenario that requires exchange of specific information. The English speaker typically plays the role of information seeker and "drives" the majority of conversations. These large-vocabulary conversations are spontaneous and free-form, with few restrictions. This collection consists of a variety of domains including force protection (e.g. checkpoint, reconnaissance, patrol), medical diagnosis and 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 [1] based on the heuristic approach of [2]. A 4-gram target LM was trained on all English transcriptions. Our phrase-based decoder is similar to Moses [3] 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 [4] on a held-out development set (≈11,000 sentence pairs) using BLEU as the tuning metric. Finally, we evaluated translation performance on a separate, unseen test set (≈9,300 sentence pairs). Most of these conversations between bilingual speakers are mediated through a human interpreter.

Of the 773K training sentence pairs, about 267K originate in ≈3,000 marked-up bilingual conversations. We use this subset to construct an anticipatory corpus for the adaptation experiments. These sentence pairs are assigned a unique utterance ID. All other sentence pairs are assigned to a default utterance ID, which signals the absence of the anticipatory relevance feature for phrase pairs derived from these instances.

4. Experimental Results

We constructed an English-Iraqi APCs from input-response pairs in the training conversations. For each source language input turn in the held-out development and test sets, we performed cross-lingual retrieval on the APC to obtain a bias corpus of potential responses in the target language. We performed retrieval in two configurations: (a) using reference transcriptions of all utterances in both languages; and (b) using ASR transcriptions (both for retrieval and translation) in both languages. The latter configuration degrades performance noticeably, but it matches the conditions of a live deployment. In the Iraqi-English experiments, we test values of the relevance list size $n \in \{50, 100, 500\}$.

The Iraqi ASR transcriptions were generated using a two-pass HMM-based system, which delivered a word error rate (WER) of 20.2% on the test set utterances. The English ASR system, which was used to transcribe the counterpart's utterances had a WER of 10.6%.

The held-out development conversations were used to tune the size of the bias corpus (i.e the number of retrieved response turns), as well as the model weights in the log-linear translation model. Tuning was performed using reference transcriptions of the Iraqi turn. The optimal settings were then used to decode the unseen test conversations for both reference transcriptions and ASR transcriptions.

Reference Transcriptions			
System	BLEU↑	TER↓	METEOR↑
Baseline	31.62	53.32	63.59
n=50	31.73*	53.11*	63.67
n=100	**31.82***	53.03*	**63.75**
n=500	31.80*	**53.00***	**63.75**

ASR Transcriptions			
System	BLEU↑	TER↓	METEOR↑
Baseline	26.93	60.38	58.20
n=50	26.98	60.12	58.21
n=100	**27.11***	60.16*	**58.26**
n=500	27.01	**60.06***	58.25

Table 3: Translation results on the test sets. Asterisked results are significantly better than the baseline ($p \leq 0.05$) using 1,000 iterations of paired bootstrap re-sampling [5]. Best results for each metric are marked in boldface.

Table 3 summarizes the translation performance of the test sets in BLEU [6], TER [7] and METEOR [8]. Results are presented for three configurations of n: 50, 100 and 500. We note that our proposed anticipatory adaptation approach outperforms the baseline across multiple metrics, both reference transcriptions and ASR transcriptions. In many instances, the differences are statistically significant. The adapted system with 100 retrieval responses (n=100) is the best scoring system for that test set.

In Table 4 we show example utterances where our adaptation approach generates better translation choices. In these examples, the conversational counterpart's utterance guides the retrieval towards contextually relevant matches, which influence lexical (hence, phrasal) selection (e.g. 'flight of stairs' vs. 'stairs' in a conversation about a corridor). Retrieval-based adaptation can also go awry, as the fourth example shows. In this example, the brevity of the preceding English turn leads to imprecise retrieval and an unreliable bias corpus, which then prefers an incorrect translation for incidental reasons.

We also compared smoothed, sentence-level BLEU scores,[2] and observed that the the n=100 adapted system scores higher than the baseline 884 times and lower than the baseline 763 times.[3] We take this as further evidence that the retrieval-based adaptation leads to small but systematic improvements in translation quality.

5. Relation to Prior Work

Online model adaptation for SMT has become an active area of research in recent years. The predominant approach is to divide the training data into discrete partitions representing either *domains* or *genres* to be adapted to [9, 10] or other linguistic phenomena of interest, such as whether the current utterance is a *question* [11]. At run-time, the domain, genre or other inferred properties of the current utterance are used to prefer phrase translation rules that originate in appropriate training data. By contrast, our approach makes no assumptions about the nature of the training data, and therefore requires no hard decisions about training set partitions and no labor-intensive manual annotation. Instead, we directly retrieve exemplars from the training set using lexical cues in order to guide the anticipatory inference.

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 [12], 'biTAM' [13] or latent dirichlet allocation [14, 15, 16]. As it also avoids data set partitioning and explicit annotation, our work is in the same spirit as these, but we do not explicitly model topic distributions.

In our previous work [16], we *incrementally* accumulated conversational history to compute a topic distribution vector. The phrasal translation rules were scored using the maximum similarity of the current conversational topic vector to all of the training conversation topic vectors from which that phrasal rule was drawn. This work is also incremental, but in contrast uses only the previous utterance of the conversational counterpart to retrieve exemplars for similarity comparisons. Here, we score phrasal rules using the maximum similarity of all of the retrieved sentences to any of the sentences from which the phrase pair was drawn.

6. Discussion and Future Directions

Conversational spoken language translation systems offer rich contextual cues that can be used to improve the MT performance. This in turn results in more usable, higher quality CSLT systems that are better able to accomplish cross-lingual communication goals in a way that is tailored to the conversation at hand. In this paper, we described a novel, turn-level anticipatory translation model adaptation technique where one participant's turn is used to anticipate,

[2] As computed by the NIST BLEU script.

[3] Of the remaining 7,662 utterances, the two systems differ in their translations of 1,867, even though their BLEU scores do not differ.

Previous Eng Turn	but his temperature how has he been hotter than normal
Baseline	*his temperature sometimes and his body is very hot*
Adaptive	*his temperature goes up sometimes and his body is very hot*
Reference	his temperature sometimes goes up and his body becomes very hot
Previous Eng Turn	can you see this corridor in front of you
Baseline	*this is the end there are stairs*
Adaptive	*at the end of it there is a flight of stairs*
Reference	at the end of it there's a staircase
Previous Eng Turn	if you can't stop it then that is an emergency situation
Baseline	*of course they call it the pressure the direct pressure on the wound or continuous*
Adaptive	*of course they call it direct pressure or continuous pressure on the wound*
Reference	of course they call it direct pressure or continuous pressure on the wound
Previous Eng Turn	good
Baseline	*personally because he is supposed to*
Adaptive	*personally because he is the foundation*
Reference	to him personally because he is the one concerned [...with the matter]

Table 4: Examples of Iraqi-to-English translations where anticipatory adaptation influences the lexical choice.

and thereby more accurately translate, the other participant's response.

The proposed approach used cross-lingual retrieval on an "anticipatory parallel corpus" of target language turns and corresponding source language responses to obtain the most relevant responses to a query turn. The retrieved responses were used to bias translation options in the translation model for the subsequent response turn in an Iraqi-Arabic-to-English translation system. We observed statistically significant improvements in translation results for most of the testing conditions, which included both reference and ASR transcripts of the bilingual test conversations. We also showed examples where the proposed approach produced better translations than the baseline system.

In this paper, we demonstrated the usefulness of turn-level context of bilingual conversations for improving MT performance. Our next goal is to develop a framework for integration of fine-grained turn-level translation model adaptation with more coarse-grained, globally driven approaches such as topic-based translation model adaptation, possibly in a neural-network-based translation model (such as [17]) where diverse sources of information can be combined to make more informed translation choices. We also plan to explore ways of detecting unreliable retrieval query input (e.g., short preceding conversational turns, as in Table 4) that can lead to unreliable translation biasing.

7. Acknowledgements

The authors would like to thank the anonymous reviewers for their helpful feedback. This work was funded in part by the DARPA BOLT program under contract number HR0011-12-C-0014.

8. References

[1] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, March 2003. [Online]. Available: http://dx.doi.org/10.1162/089120103321337421

[2] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation," in *NAACL-2003: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*. Morristown, NJ, USA: Association for Computational Linguistics, 2003, pp. 48–54.

[3] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ser. ACL-2007. Stroudsburg, PA, USA: Association for Computational Linguistics, 2007, pp. 177–180. [Online]. Available: http://dl.acm.org/citation.cfm?id=1557769.1557821

[4] F. J. Och, "Minimum error rate training in statistical machine translation," in *ACL-2003: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*. Morristown, NJ, USA: Association for Computational Linguistics, 2003, pp. 160–167.

[5] P. Koehn, "Statistical significance tests for machine translation evaluation," in *EMNLP*, Barcelona, Spain, July 2004, pp. 388–395.

[6] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: A method for automatic evaluation of machine translation,” in *ACL-2002: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Morristown, NJ, USA: Association for Computational Linguistics, 2001, pp. 311–318.

[7] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, “A study of translation edit rate with targeted human annotation,” in *Proceedings AMTA*, August 2006, pp. 223–231. [Online]. Available: http://www.mt-archive.info/AMTA-2006-Snover.pdf

[8] S. Banerjee and A. Lavie, “METEOR: An automatic metric for MT evaluation with improved correlation with human judgments,” in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ann Arbor, Michigan: Association for Computational Linguistics, June 2005, pp. 65–72. [Online]. Available: http://www.aclweb.org/anthology/W/W05/W05-0909

[9] G. Foster and R. Kuhn, “Mixture-model adaptation for SMT,” in *Proceedings of the Second Workshop on Statistical Machine Translation*, ser. StatMT-2007. Stroudsburg, PA, USA: Association for Computational Linguistics, 2007, pp. 128–135. [Online]. Available: http://dl.acm.org/citation.cfm?id=1626355.1626372

[10] S. Matsoukas, A.-V. I. Rosti, and B. Zhang, “Discriminative corpus weight estimation for machine translation,” in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, ser. EMNLP-2009. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 708–717. [Online]. Available: http://dl.acm.org/citation.cfm?id=1699571.1699605

[11] A. Finch and E. Sumita, “Dynamic model interpolation for statistical machine translation,” in *Proceedings of the Third Workshop on Statistical Machine Translation*, ser. StatMT-2008. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008, pp. 208–215. [Online]. Available: http://dl.acm.org/citation.cfm?id=1626394.1626428

[12] Y.-C. Tam, I. Lane, and T. Schultz, “Bilingual LSA-based adaptation for statistical machine translation,” *Machine Translation*, vol. 21, no. 4, pp. 187–207, Dec. 2007. [Online]. Available: http://dx.doi.org/10.1007/s10590-008-9045-2

[13] B. Zhao and E. P. Xing, “BiTAM: Bilingual topic admixture models for word alignment,” in *In Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL ‘06)*, 2006.

[14] Z. Gong, Y. Zhang, and G. Zhou, “Statistical machine translation based on LDA,” in *Universal Communication Symposium (IUCS), 2010 4th International*, 2010, pp. 286–290.

[15] V. Eidelman, J. Boyd-Graber, and P. Resnik, “Topic models for dynamic translation model adaptation,” in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ser. ACL-2012. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, pp. 115–119. [Online]. Available: http://dl.acm.org/citation.cfm?id=2390665.2390694

[16] S. Hewavitharana, D. N. Mehay, S. Ananthakrishnan, and P. Natarajan, “Incremental Topic-Based Translation Model Adaptation for Conversational Spoken Language Translation.” in *ACL-2013: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, 2013, pp. 697–701.

[17] J. Devlin, R. Zbib, Z. Huang, T. Lamar, R. Schwartz, and J. Makhoul, “Fast and robust neural network joint models for statistical machine translation,” in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2014, pp. 1370–1380.