Rapid Development of Speech Translation using Consecutive Interpretation

Matthias Paulik and Alex Waibel

Interactive Systems Laboratories (interACT)
Carnegie Mellon University, USA and Karlsruhe Institute of Technology, Germany
{paulik, waibel}@cs.cmu.edu

Abstract

The development of a speech translation (ST) system is costly, largely because it is expensive to collect parallel data. A new language pair is typically only considered in the aftermath of an international crisis that incurs a major need of cross-lingual communication. Urgency justifies the deployment of interpreters while data is being collected. In recent work, we have shown that audio recordings of interpreter-mediated communication can present a low-cost data resource for the rapid development of automatic text and speech translation. However, our previous experiments remain limited to English/Spanish simultaneous interpretation. In this work, we examine our approaches for exploiting interpretation audio as translation model training data in the context of English/Pashto consecutive interpretation. We show that our previously made findings remain valid, despite the more complex language pair and the additional challenges introduced by the strong resource-limitations of Pashto.
Index Terms: speech translation, machine translation, parallel speech

1. Introduction

The rapid development of speech translation (ST) systems for a new language pair or domain often fails due to the time-consuming and costly effort of acquiring sufficient amounts of suitable training data. In fact, the prohibitively high costs attached to training data acquisition are one main reason why the development of deployable ST systems remains limited to only a handful of languages. New language pairs are typically only considered for ST development after a major need for cross-lingual verbal communication just arose—justifying the high development costs. In these situations, communication has to rely on interpreters until suitable systems become available. We examine the feasibility of using audio recordings of consecutive interpretation (CI) as a novel, low-cost training data resource for translation model (TM) training. In other words, we aim to learn automatic *translation* from *interpretation* audio.

In recent work [1, 2] we proposed the use of automatically transcribed audio recordings of interpreter-mediated communication scenarios for training statistical translation models (TMs). We refer to such audio recordings as 'parallel speech' (pSp) audio. In [1], we reported that the training corpus size-dependent performance of pSp-trained TMs basically mirrors the training corpus size-dependent performance of TMs trained on parallel text, just at a lower level. Our results suggested that more (in-domain) training data results in both cases in an improved translation performance, while successively higher amounts of training data are necessary to achieve the same improvements in BLEU. Further, we observed that TMs trained exclusively on n interpreted words achieve a similar text translation performance, measured in BLEU, as TMs trained on $n \cdot 10^{-1}$ translated words. In [2] we further reported statistically significant improvements in BLEU by increasing a parallel text corpus of 100k manually translated words with 752k interpreted words, stemming from the automatic transcription of 92h of pSp audio. Our previous experiments remain limited to pSp audio of simultaneous interpretation (SI), as provided during sessions of the European Parliament, between the rather simple[1] language pair English/Spanish. In this work, we examine if the reported findings remain valid in the context of CI between English (En) and the under-resourced language Pashto (Pa).

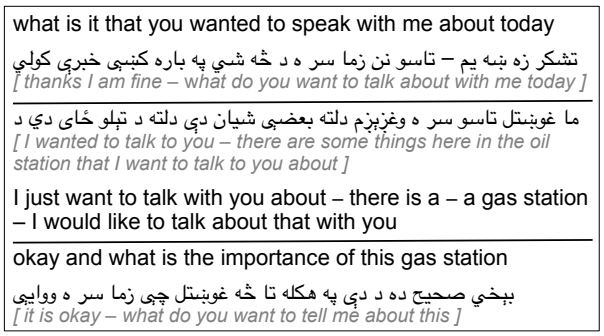

Figure 1: Consecutive interpretation example.

The remainder of this paper is organized as follows. In Section 2 we shortly highlight the challenges faced when exploiting (consecutive) interpretation audio for TM training. Section 3 describes our experimental setup and the general approach. In Section 4, we examine the situation where only pSp audio but no parallel text data is available for ST development. Section 5 reports our results for exploiting pSp audio as a training resource in addition to parallel text. Finally, in Section 6 we summarize our results and briefly discuss their significance in the context of ST development.

2. Challenges

One major challenge faced when training TMs from interpretation is the significant difference between translation and interpretation, as explained in the following.

Two basic forms of interpretation can be distinguished.

[1]Simple in terms of 'complexity' for machine translation, which is influenced by many factors, as for example amount of previous research, morphological richness of the involved languages, available data-resources, word re-orderings, etc.

Copyright © 2010 ISCA

	Native		Interpr.	
	En	Pa	En	Pa
audio [h]	23.0	25.2	26.7	29.5
words [k]	358	374	333	399

Table 1: Parallel speech audio statistics.

	Pa→En	
	Dev	Eval
audio [min]	45.8	24.0
words [k]	6.7	3.6

Table 2: Development and evaluation set statistics.

In simultaneous interpretation, the interpreter renders the interpretation simultaneously, while the source speaker continuously speaks. In consecutive interpretation, source speaker and interpreter take turns, resulting in less severe time constraints for the interpreter. Due to these less severe time constraints, CI exhibits "more accurate, equivalent, and complete interpretations" than SI [3]. However, both forms of interpretation are cognitively very demanding tasks, that can only be accomplished by applying special interpretation strategies. The interpretation strategy of 'dropping form' can be identified as one of the main reasons why interpretation and translation differ strongly. Dropping form means that interpreters immediately and deliberately discard the wording and retention of the mental representation of the message [4]. Only by discarding the words, sentence structure, etc., interpreters—in SI as well as in CI—are able to concentrate on the meaning of the message and its reformulation in the target language [3]. The reason for this lies within the limitations of the human short-term memory. Only up to six or seven items can be retained in short-term memory, and only if we give all of our attention to them [5].
Further differences between interpretation and translation result from the fact that "interpreters also elaborate and change information and they do not only convey all elements of meaning, but also the intentions and feelings of the source speaker" [6]. We speculate that the latter effect is more prevalent in CI than in SI, as CI scenarios tend to be more personal and the interpreter has more time to elaborate. The En/Pa CI dialog shown in Figure 1 gives an example for some of the significant differences between interpretation and translation. Each native speech utterance is accompanied by its CI utterance in the example. Further, a manual translation of the non-English parallel speech is provided.

We argue that pSp audio is of special interest to ST development in the context of under-resourced languages, where in-domain parallel text data is especially hard to come by. As we intend to automatically transcribe pSp audio using automatic speech recognition (ASR) for a cost-effective use, we face another major challenge; a potentially high word error rate (WER) of the under-resourced ASR system.

3. Experimental Setup

3.1. Data Resources and Scoring

Our experiments are based on data resources provided within US Darpa's TransTac project. TransTac aims to rapidly develop ST for real-world tactical situations. Typical scenarios are in the form of interviews, where an English-speaking soldier interviews for example a Pashto-speaking Afghani, compare also Figure 1. Only very limited amounts of data resources are available for En/Pa ST development. Table 1 lists the statistics of the En/Pa pSp corpus. It shows the amount of native speech (En interviewer, Pa respondent) and interpreter speech in hours of audio and number of uttered words. For each utterance in the pSp corpus, we have manual reference transcriptions and manual reference translations available. In addition to the pSp corpus, we use a 'traditional' En/Pa parallel text corpus of manual translations. This corpus has 12.4k translated Pashto respondent utterances. The Pashto part comprises 260k words; the English part has 214k words.

Table 2 lists the statistics of the Pa→En development (dev) and evaluation (eval) set. Both sets are based on native Pashto respondent speech and feature only one reference *translation* for BLEU score computation.

3.2. ASR Systems

The employed ASR systems are developed with the Janus Recognition Toolkit (JRTk), featuring the IBIS single pass decoder [7]. The SRI Language Model Toolkit [8] is used for language model (LM) training. Both, English and Pashto ASR, feature only one decoding pass with incremental, unsupervised feature space adaptation (constrained maximum likelihood linear regression). The systems are tuned to the real-time requirements of TransTac. Acoustic model (AM) training involves in both cases several iterations of standard Viterbi training. For the English system, we also apply several iterations of feature space adaptive (FSA) Viterbi training, followed by several iterations of FSA boosted maximum mutual information training [9]. The English AM is estimated on approximately 83.5h of TransTac data from previous phases of the project, including native speech and interpreter speech, and 34.4h of broadcast new data. The English 4-gram LM is estimated on approximately 74.8M running words. LM training data includes the transcriptions used for AM training as well as web data. The Pashto AM is estimated on the 25.2h of Pashto respondent speech included in our pSp corpus. For LM training (3-gram LM), we rely on the manual transcription of this respondent speech, which amounts to 374k words. Performance numbers for both ASR systems are given at the beginning of Section 4.

3.3. Sentence Alignment

In order to utilize the En/Pa pSp audio corpus in a standard TM training setup, we have to create a sentence-aligned bilingual text corpus first. English and Pashto ASR provides the necessary transcriptions. For sentence alignment, we can exploit the fact that each speaker takes turns in CI, with each speaker producing only a few utterances in each turn. To introduce speaker-turn-based sentence alignment, we rely on manual utterance segmentation and manual speaker ids[2]. All of our training runs are based on aligned speaker turns, even when manual translations are used for model building. This is possible, since each speech utterance is accompanied with a manual translation in the corpus. Our decoding/scoring runs on dev and eval observe the manual speech utterance segmentation.

[2] As interpreter and interviewer/respondent are recorded on different audio channels, we argue that an automatic utterance segmentation and speaker identification will provide very similar performance.

	Native		Interpr.	
	En	Pa	En	Pa
PPL	68	-	75	196
WER [%]	16.3	-	30.7	44.9

Table 3: Parallel speech audio: PPL and WER

3.4. TM Training Setup and MT Decoder

Our standard TM training setup extracts phrase tables from the created bilingual training corpus by using the GIZA++ toolkit [10] in combination with University Edinburgh's training scripts, as provided during the NAACL 2006 Workshop on Statistical Machine Translation [11]. The GIZA++ toolkit is run with its standard parameter settings.

For MT, we use the Interactive Systems Labs beam search decoder [12]. The decoder combines multiple model scores to find the best translation:

- The translation model.
- A 4-gram target language model. The applied English LM is identical to the LM used for ASR.
- A word reordering model that assigns higher costs to longer distance reordering. We use a reordering window of 4.
- Simple word and phrase count models.

To optimize the system parameters, we use Minimum Error Rate (MER) Training as described in [13].

4. CI Audio as Only Data Source

In a situation where only untranscribed pSp audio of CI is available, the minimal requirement for ST development are two ASR systems to enable the automatic transcription of source and target language speech. In the case of ST development between a resource-rich and a under-resourced language, ASR systems for the resource-rich language may already be available. In our case, we have an in-domain English ASR system from previous phases of the TransTac project available, as previous phases considered ST between (a) English and (b) Iraqi, Farsi and Dari. However, we have no pre-existing Pashto ASR on hand. To enable Pa→En speech translation and to be able to automatically transcribe additional pSp audio, we train a Pashto ASR system on the 25.2h of Pashto respondent speech found in our pSp corpus. For AM and LM training, we rely on the manual transcription of this respondent speech (374k words). Table 3 lists the English and Pashto WER and LM perplexity for the automatically transcribed parts of the pSp corpus. The interpreter speech frequently suffers from a heavy foreign accent, explaining the significantly higher WER on interpreter speech compared to native speech. The Pashto WER on the Pa→En development and test set is 33.7% and 33.9%, respectively. The LM perplexity is 157 and 148, respectively.

To examine if our hypotheses made in [1] regarding the performance of pSp-trained TMs remain valid in the context of En/Pa CI, we examine three different systems. System A uses TMs trained on the manually transcribed and translated Pashto respondent speech that is present in our pSp corpus. In System B the English translations are replaced by the manual transcription of the interpreter speech. System C finally uses the automatic transcription (30.7% WER) of English interpreter

	text		speech	
	dev	eval	dev	eval
A	17.6	17.8	14.6	15.2
B	11.8	13.0	10.5	10.0
C	10.9	10.5	9.4	10.2

Table 4: Pa→En translation performance.

	token	type
A	98.8	92.9
B	98.1	90.0
C	98.1	90.3
D	97.8	88.6
D+C+F	99.1	95.1

Table 5: Vocabulary and corpus coverage.

speech. While system C does not suffer from word errors on the Pashto side (we use here the Pashto reference transcription since we trained the Pashto ASR on these transcriptions), the English WER is on the same level as the worst WER level considered in [1]. Table 4 lists the text and speech translation performance in BLEU for all three systems. Table 5 lists the English type and token coverage of the training corpora A and B in regard to dev, showing that corpus coverage does not play an important role. As we expect system B and C to perform on the same level as a system that is trained on approximately 40k manually translated words, we compute the corpus size-dependent text translation performance of system A for increments of 10k words, until system A meets the performance of system B. The result is depicted in Figure 2. It shows that the prediction was accurate. Figure 3 compares the corpus size-dependent text translation performance of system A and B in increments of 90k words. We observe the same trend as described in [1].

5. CI Audio as Additional Data Source

To further examine the value of pSp audio as TM training data in addition to parallel text, we estimate a TM on the parallel text corpus of 260k translated Pashto words. We refer to the system using this TM as system D. We then increase the parallel text corpus with the training corpus of system A, B or C and estimate new TMs, resulting in systems D+A, D+B and D+C. Table 6 gives an overview of the text and speech translation performance of these systems.

With English and Pashto ASR available, it is possible to automatically transcribe more pSp audio, promising further gains in translation performance at a relatively low cost. For example, we can automatically transcribe the part of the pSp corpus formed by English interviewer speech (16.3% WER) and re-

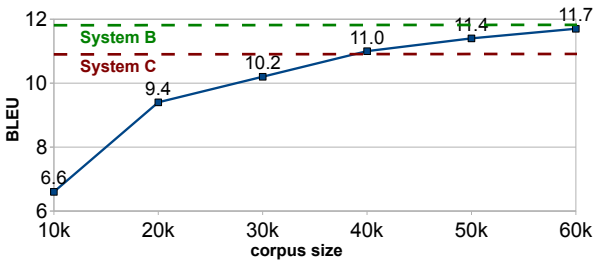

Figure 2: BLEU development, system A

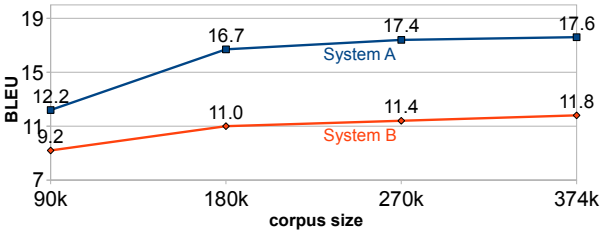

Figure 3: BLEU development, system A & B

	text		speech	
	dev	eval	dev	eval
D	**12.3**	**12.3**	**11.2**	**10.0**
D+A	18.4	17.5	16.0	14.2
D+B	14.6	14.7	12.7	12.2
D+C	13.8	13.4	11.6	12.0
D+C+F	**14.7**	**14.9**	**12.5**	**12.4**

Table 6: Parallel text plus pSp audio

spective Pashto interpretation (44.9% WER)—referred to in the following as training data F. Despite the very high Pashto WER, we achieve further gains in text and speech translation performance by adding training data F to D+C, as shown in the last row of Table 6. The observed improvements for system D+C+F compared to system D are statistically significant ($p < 0.05$). These results are achieved by weighting training data D+C and training data F differently. In the case of text (speech) translation, D+C was repeated 3 (4) times in the final training corpus D+C+F.

6. Results and Discussion

Our results show that our previous findings [1, 2] regarding pSp trained translation models, made in the context of English/Spanish simultaneous interpretation, remain valid in the context of consecutive interpretation between English and the resource-limited language Pashto. We have shown that training data in the form of automatically transcribed pSp audio of CI can (a) replace parallel text for TM training; and (b) that traditionally trained TMs can be improved with such training data. These results further support our hypothesis that automatically transcribed pSp audio (of CI as well as SI) can present a low-cost data resource that is valuable for rapid development of automatic text and speech translation systems.

Compared to our previous findings, we observe a similar or even slightly better yield of pSp audio compared to parallel text, despite the more complex language pair English/Pashto and despite higher word error rates. This result indicates that pSp audio of CI may yield better automatic translation performance than pSp audio of SI. The more 'complete' interpretations (compare Section 2) of CI, in addition to the for CI less complex task of sentence alignment, support this hypothesis.

We did not apply any word-confidence based filtering of ASR hypotheses before TM training. Future approaches that automatically identify (interpretation) ASR hypotheses that are problematic in terms of WER or content could be of special interest in the context of pSp audio. We further believe that future work has to address larger amounts of pSp audio (of SI and CI alike) and more language pairs, to further support hypotheses made regarding the translation performance of pSp-trained TMs. While the attached collection effort of additional pSp audio can be considered the biggest obstacle, one has to realize that a) interpretation happens daily on a massive scale, b) simultaneous interpretation typically involves considerable amounts of equipment (sound proof booths, etc.) that directly enable the recording of pSp audio and c) that huge amounts of money flow into the development of ST systems for CI like situations. The latter point implies that there are many CI situations in which the recording of source and target language speech is feasible. Therefore, our results promise substantial improvements in automatic translation of text and speech, achieved at a relatively low additional cost, by collecting more pSp audio.

7. Acknowledgements

This work is in part supported by the US DARPA under the TransTac program. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

8. References

[1] M. Paulik and A. Waibel, "Automatic Translation from Parallel Speech: Simultaneous Interpretation as MT Training Data," in *ASRU*, Merano, Italy, 2009.

[2] ——, "Spoken Language Translation from Parallel Speech Audio: Simultaneous Interpretation as SLT Training Data," in *ICASSP*, Dallas, TX, USA, 2010.

[3] R. Santiago, "Consecutive Interpreting: A Brief Review," Accesed on January 11, 2010 at http://home.earthlink.net/∼terperto/id16.html, 2004.

[4] D. Seleskovitch, *Interpreting for International Conferences*. Pen and Booth, Arlington, VA, 1978.

[5] F. Smith, *Reading Without Nonsense*. NY Teachers College Press, NY, NY, 1985.

[6] K. Kohn and S. Kalina, "The Strategic Dimension of Interpreting," *Meta: Journal des traducteurs*, vol. 41(1), pp. 118–138, 1996.

[7] H. Soltau, F. Metze, C. Fügen, and A. Waibel, "A One Pass-decoder Based on Polymorphic Linguistic Context Assignment," in *ASRU*, Madonna di Campiglio Trento, Italy, December 2001.

[8] A. Stolcke, "SRILM – an extensible language modeling toolkit," in *Intl. Conf. on Spoken Language Processing*, Denver, CO, USA, September 2002.

[9] D. Povey, D. Kanevsky, B. Kingsbury, and B. Ramabhadran, "Boosted MMI for model and feature-space discriminative training," in *ICASSP*, Las Vegas, LV, USA, April 2008.

[10] F. Och and H. Ney, "A Systematic Comparison of Various Statistical Alignment Models," *Computational Linguistics*, vol. 29(1), pp. 19–51, 2003.

[11] P. Koehn and C. Monz, "Manual and Automatic Evaluation of Machine Translation between European Languages," in *Proc. on the Workshop on Statistical Machine Translation*, New York City, USA, 2006, pp. 102–121.

[12] S. Vogel, "SMT Decoder Dissected: Word Reordering," in *Proc. of Coling*, Beijing, China, 2003.

[13] F. Och, "Minimum Error Rate Training in Statistical Machine Translation," in *Proc of ACL*, Sapporo, Japan, 2003.