SPOKEN LANGUAGE TRANSLATION FROM PARALLEL SPEECH AUDIO: SIMULTANEOUS INTERPRETATION AS SLT TRAINING DATA

Matthias Paulik and Alex Waibel

Carnegie Mellon University (USA) and Universität Karlsruhe (Germany)
{paulik, waibel}@cs.cmu.edu

ABSTRACT

In recent work, we proposed an alternative to parallel text as translation model (TM) training data: audio recordings of parallel speech (pSp), as it occurs in any communication scenario where interpreters are involved. Although interpretation compares poorly to translation, we reported surprisingly strong translation results for systems based on pSp trained TMs. This work extends the use of pSp as a data source for unsupervised training of all major models involved in statistical spoken language translation. We consider the scenario of speech translation between a resource rich and a resource-deficient language. Our seed models are based on 10h of transcribed audio and parallel text comprised of 100k translated words. With the help of 92h of untranscribed pSp audio, and by taking advantage of the redundancy inherent to pSp (the same information is given twice, in two languages), we report significant improvements for the resource-deficient acoustic, language and translation models.

Index Terms— SLT, ASR, MT, parallel speech

1. INTRODUCTION

In [1] we demonstrated that statistical translation models can be trained in a fully automatic, unsupervised manner from audio recordings of human interpretation scenarios. We used automatic speech recognition (ASR) to create a bilingual parallel translation model (TM) training corpus from the parallel speech (pSp) audio of source language speaker and simultaneous interpreter. Even when manually transcribed (0% word error rate), interpretation (parallel speech) differs significantly from translation (parallel text). In simultaneous interpretation, this strong difference stems mostly from the anticipationary and compensatory strategies interpreters have to apply to keep pace with the source language speaker. Despite the mismatch between interpretation and translation, we reported in [1] surprisingly strong text translation and speech translation results for our parallel speech audio trained translation models.

In this work, we extend the use of pSp audio as a data source for unsupervised training of all major models involved in statistical spoken language translation (SLT); ASR acoustic model (AM) and ASR language model (LM) as well as machine translation (MT) translation model and MT target LM. Specifically, we explore techniques for unsupervised AM and LM training. Further, we exploit the parallel nature of pSp audio to train translation models and to introduce light supervision for SLT model training. We conduct our experiments within the scenario of automatic speech translation between a resource rich language and a resource-deficient language. For the resource-deficient language, we have 10h of manually transcribed audio available as well as a parallel text corpus comprised of 100k translated words. We seek to improve the performance of the statistical models affected by the resource-deficiency.

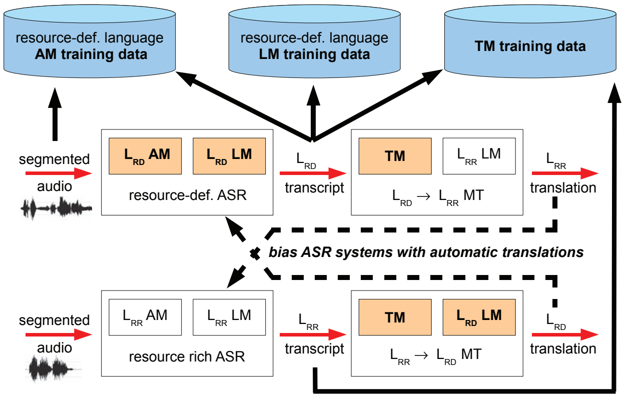

Fig. 1. Extracting SLT training data from parallel speech.

2. EXPERIMENTAL SETUP

2.1. System Architecture

In the proposed scenario of speech translation between a resource rich language L_{RR} and a resource-deficient language L_{RD}, we seek to improve the statistical SLT models that suffer from the resource-deficiency, by automatically extracting training data from pSp audio. Figure 2.1 shows our system architecture. The overall system consists of two SLT sub-systems, each featuring an ASR component and a MT component. The ASR systems accept pre-segmented speech utterances; we use a HMM based, language independent speech/non-speech audio segmentation. The models affected by the resource-deficiency are highlighted by color in the diagram. The core components necessary to extract SLT training data are the two ASR systems. Together with the input audio, automatic transcriptions for L_{RD} can be used for unsupervised AM training. The transcriptions can also be used as additional LM training data. Further, the hypotheses of both ASR systems can be tied together in a parallel training corpus suitable for TM training, as shown in [1]. Similar to previous works [2, 3, 4], we exploit the parallel information given in the respective other language audio stream to bias the ASR systems for an improved transcription performance. In the proposed context, such an improved ASR performance directly affects the quality of the extracted training data.

2.2. Data & Scoring

European Parliament Plenary Sessions (EPPS) are broadcast live via satellite in the different official languages of the European Union. Each language L_i has a dedicated audio channel. An interpreter provides the simultaneous interpretation in language L_i whenever a

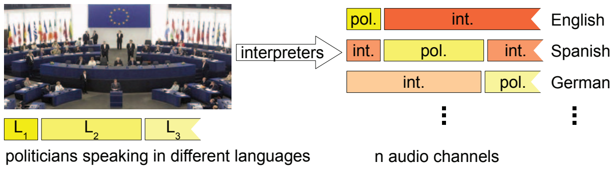

Fig. 2. EPPS live broadcast: a data source for parallel speech audio.

	English-to-Spanish			**Spanish**-to-English		
	TDev	dev	eval	TDev	dev	eval
utterances	1256	1287	1926	1589	1707	2085
words [k]	17.4	27.9	26.0	14.7	22.4	25.8
audio [h]	1.6	3.2	2.7	1.5	2.3	2.7
base WER	13.1	13.9	12.2	26.1	26.9	27.1

Table 1. Data statistics: development and evaluation sets.

politician is speaking in a language $L_{j \neq i}$. In the case that a politician is speaking in the respective language of an audio channel, the original speech of the politician is being broadcast on that channel. The language dedicated audio channels, depicted on the right hand side of Figure 2, form an excellent source for pSp audio based on simultaneous interpretation.

For our experiments we use the EPPS part of the English and Spanish TC-STAR spring 2007 verbatim task development and evaluation sets. It has to be noted that these sets are only comprised of politician speech and do not include any interpreter speech. The Spanish and English audio in the development ('dev') and evaluation ('eval') set was therefore recorded from the one audio channel shown on the left hand side of Figure 2. We do not have pSp available for dev and eval. In addition to these sets, we use one Parliamentary session (26OCT04) extracted from the TC-STAR verbatim 2005 development set. This one session, referred to in the following as training-development ('TDev') set, features parallel speech and is included in our data base of pSp audio used for model training. With the exception of TDev, we do not have any manual reference transcription available for our pSp audio corpus. The pSp corpus includes 67 sessions from the time period 08SEP05-01JUN06. The dev set has sessions from 06JUN05-06SEP05 and the eval set has sessions from 12JUN06-28SEP06. English/Spanish supervised AM training data is from the time period May 2004 to January 2005. Supervised LM training data and parallel text data is from the time period April 1996 to May 2005 (excluding TDev). Detailed data statistics for the dev, eval and training sets are shown in Table 1 and Table 2.

For scoring ASR and MT performance we use non-punctuated, lowercased references. ASR performance is measured in word error rate (WER) and MT performance is measured in IBM BLEU using two reference translations.

2.3. Baseline ASR & MT Systems

Our ASR systems are based on the Janus Recognition Toolkit (JRTk), featuring the IBIS single pass decoder. For LM training, we use the SRI Language Model Toolkit [5].

The English ASR system produces word error rates in the range of 12-14% on our data sets; compare Table 1. A detailed system description can be found in [1]. The Spanish ASR system, based on sub-phonetically tied three-state HMMs, features a single, speaker independent decoding pass. The AM is trained on 10h Spanish EPPS data via three iterations of Viterbi training. The 3-gram LM is estimated on 179.6k running words from the AM training data reference transcriptions and the Spanish side of the parallel text corpus used for supervised TM training. In order to avoid high out-of-vocabulary rates, we use a large recognition dictionary with 74.2K pronunciation entries. This resource-limited Spanish ASR system yields WERs in the range of 26-27% on our data sets; compare Table 1.

For MT, we use the ISL beam search decoder [6]. To optimize the system towards a maximal BLEU score, we apply MER training as described in [7]. Our TMs consist of phrase-to-phrase translation pairs that we extract from a bilingual text corpus with the help of the GIZA++ toolkit [8] and University Edinburgh's phrase model training scripts. The MT target LMs are identical to the respective ASR LMs.

	transcriptions	parallel text	pSp
sent. / utt. [k]	6.5	3.9	52.3
words [k]	79.6	100.0	751.8
audio [h]	10.0	N/A	91.7

Table 2. Data statistics: Spanish SLT training data.

3. PSP AUDIO FOR ASR MODEL TRAINING

Unsupervised AM training relies on automatic transcriptions created with an initial ASR system. The success of unsupervised AM training usually depends strongly on the ability to exclude erroneous transcriptions from training. The common approach is to use word confidences for selecting transcriptions suitable for training. Lightly supervised AM training [9] refers to the case where some imperfect human transcriptions, for example closed-captions provided during television broadcasts, can be used to either bias the initial ASR system for an improved transcription performance or to filter erroneous ASR hypotheses. In this work, we examine unsupervised AM training and lightly supervised AM training. We introduce light supervision with the help of pSp audio of simultaneous interpreters, as proposed in [10].

To introduce light supervision based on English pSp audio for Spanish AM and LM training, we automatically translate the English parallel speech into Spanish and bias the Spanish ASR LM to prefer n-grams seen in the automatic translation. We distinguish between two different types of LM bias; a 'session bias' and an 'utterance bias'. Session bias refers to the case where we first automatically translate the English audio of one complete European Parliament session into Spanish, and we then interpolate the baseline Spanish LM with a LM build on the automatic translation. Utterance bias, on the other hand, refers to the case where we bias the Spanish LM for each Spanish speech utterance. We achieve this by first translating the English speech snippet that starts/ends 6 seconds before/after the Spanish utterance starts/ends. We then prefer the uni-grams found in the translated speech snippet, by boosting the baseline Spanish LM probability of these uni-grams, similar to a cache language model. The boosting of the uni-gram probability is realized by subtracting a discount value d from the (positive) LM log score of the current ASR hypothesis. The discount value d for a uni-gram u is estimated as follows:

$$d(u) = \begin{cases} w * LM_{score}(u) & \text{for } LM_{score}(u) \geq t \\ 0 & \text{for } LM_{score}(u) < t \end{cases}$$

with $LM_{score}(u)$ being the baseline LM score for the uni-gram u and weight w and threshold t estimated on TDev via a grid search.

baseline	session bias	session & utt. bias
26.1	25.4	24.5

Table 3. Biasing ASR with pSp; WER on TDev.

Table 3 shows the influence of the session LM bias and the combination of utterance LM bias and session LM bias on the Spanish WER on TDev; the WER is reduced by 6% relative from 26.1% to 24.5%.

For unsupervised and lightly supervised AM training, we utilize ASR word confidences in the following manner: speech frames associated with words that have an ASR word confidence of $c < 0.8$ are ignored; all other speech frames contribute to the training with a weight of 1. The value of c was estimated on our dev set. Training itself is realized via three iterations of Viterbi training. All iterations include 10h of manually transcribed audio plus 92h of automatically transcribed audio. Results obtained with the re-trained AMs are listed in Table 5, along with results for unsupervised LM training. The first two columns of Table 5 specify if the baseline AM/LM was used or a model trained with additional 92h of automatically transcribed Spanish speech. The case of a light supervision during ASR decoding via a session+utterance bias is marked with a subscript $_b$. For example, the last row in the table refers to the case where we used the biased baseline ASR system to create additional AM and LM training data. The values shown in brackets represent the WER on TDev, when biasing the ASR with knowledge from the English parallel speech. Since we do not have English pSp available for dev and eval, such a bias is not possible on these data sets. The results show that light supervision during training benefits the ASR performance.

LM	TDev	dev	eval
base	182	269	276
+92h	129*	202	206
+92h$_b$	127*	200	204

Table 4. LM training: perplexity.

In contrast to AM training, we did not utilize ASR word confidences during LM training. We estimated a LM on the Spanish ASR first-best hypotheses and interpolated this LM with the baseline LM. The interpolation weight was chosen to minimize the LM perplexity (PPL) on the dev set. Table 4 lists the PPL of the baseline LM and of the interpolated LMs, using transcriptions from the baseline and biased baseline Spanish ASR during training. The LM used to compute the TDev PPLs (marked by *) did not include automatic transcriptions of TDev itself. We found that, while the PPL decreases much stronger if ASR first best hypotheses of the same session are included in the LM, ASR transcription performance does not benefit due to an overly strong bias towards transcription errors made by the initial ASR. Therefore, whenever we automatically transcribe our pSp corpus with an ASR system that includes a re-trained LM, we use session specific LMs that do not include ASR transcripts of the very same session.

4. PSP AUDIO FOR MT MODEL TRAINING

In [1] we demonstrated how statistical TMs can be trained solely from pSp audio, without having any traditional MT training corpora of sentence aligned, bilingual translations available. In this work, we examine how translation models trained on a small amount of traditional MT training data can benefit best from pSp audio. Following results from [1], we align the bilingual parallel speech ASR transcripts by exploiting the time alignment that is inherently given for simultaneous interpretation. Since the interpreter has to keep pace with the source language speaker, the target language interpretation that is related to a specific source utterance, occurs at approximately the same time. Similar to the target audio snippet used to bias the ASR LM, we align to each Spanish (English) utterance the ±2 seconds padded English (Spanish) audio snippet. The aligned audio is transcribed with the English and Spanish ASR systems and then added to the supervised TM training corpus. We examined a higher weighting of word alignments that stem from the supervised part of the combined corpus to aid the GIZA++ word alignment process on the pSp part. We achieve this higher weighting by simply duplicating the supervised training corpus x times. Figure 3 gives an overview of the Sp-to-En text translation (0% Spanish WER) results on the dev set for $x \in [1-8]$. The figure also list translation performance numbers in BLEU for the baseline TM, trained only on supervised training data, and for a TM trained only on the automatically transcribed and ±2 seconds aligned pSp corpus. The best translation results are achieved by adding the supervised parallel text corpus of manual translations 5 times to the combined training corpus. Tables 6 and 7 list the in this manner achieved text translation results for both translation directions. It has to be noted that the presented results were obtained with ASR transcriptions created with the baseline ASR systems.

AM	LM	TDev	dev	eval
base	base	26.1 [24.5_b]	26.9	27.1
+92h	base	24.0 [23.0_b]	24.9	25.5
base	+92h	24.5 [23.3_b]	25.7	25.5
+92h	+92h	22.5 [21.5_b]	24.0	24.2
+92h$_b$	+92h$_b$	22.0 [21.6_b]	23.5	23.8

Table 5. AM & LM training: WER.

TM	TDev	dev	eval
base	41.1	35.1	35.2
+92h	44.5	37.9	37.8

Table 6. TM training: Sp-to-En text translation, BLEU score.

5. SPEECH TRANSLATION RESULTS

So far, we presented results for an improved Spanish transcription performance and an improved text translation performance for both translation directions. The presented improvements were achieved by automatically extracting additional training data from our pSp corpus and re-training the statistical ASR and MT models with this additional data. In this section, we present our results for the complete SLT chain of ASR and subsequent MT on the ASR first best

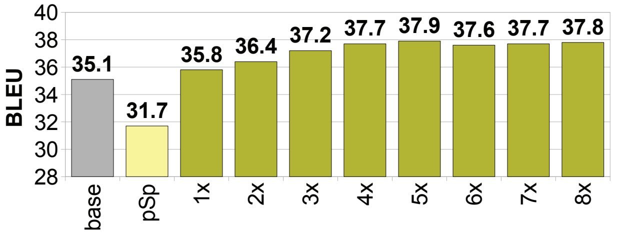

Fig. 3. Sp-to-En text translation results on dev, depending on the weight of the supervised training corpus.

LM	TM	TDev	dev	eval
base	base	26.0	27.2	25.7
+92h	base	27.9	29.1	27.5
base	+92h	27.7	28.3	27.6
+92h	+92h	30.5	30.6	29.6

Table 7. TM training: En-to-Sp text translation, BLEU score.

LM_{MT}	TM		TDev	dev	eval
base	base		24.0	22.4	21.6
+92h	+92h		28.5	25.7	25.2
+92h	92h		23.8	20.4	19.9

Table 8. En-to-Sp speech translation results in BLEU.

hypotheses. We also pay special attention to the case of a strong resource-limitation, in which only 10h of transcribed Spanish AM data is available, but no baseline MT.

Table 8 list the speech translation results for En-to-Sp. We compare results of the baseline SLT system with a SLT system that includes unsupervised training data created with the baseline Spanish ASR. The eval set BLEU score increases by 3.2 points from 21.6 to 25.2 for the re-trained SLT system. The case where no baseline automatic translation is possible due to the lack of parallel text data, is shown in the last row. With a TM trained solely on 92h of pSp audio, we achieve a translation performance of 19.9 BLEU points.

In Table 9 we show speech translation results for Sp-to-En. Here, we examine two additional scenarios: first, we examine the effect of lightly supervised AM and LM training on the speech translation end result (row 3) and second, we address the effect of the improved transcription performance on translation model training (row 5). Specifically, the results in row 5 refer to the case where the pSp automatic transcriptions used for TM training came from the already re-trained ASR. All other listed results were achieved with all models re-trained with pSp transcription that either came from the baseline ASR or the biased baseline ASR. Re-training the SLT models with baseline ASR transcripts improved the eval BLEU score by 3.0 points from 25.3 to 28.3. Using ASR hypotheses from the biased Spanish ASR did not improve the overall speech translation result on our evaluation set, although ASR transcription performance is slightly improved, as shown in Section 4. In the scenario where no parallel text data for TM training is available, we achieve an eval BLEU score of 24.9 — only slightly below the translation performance of the baseline system that is based on parallel text data. The translation performance of the pSp only system can be further increased by 0.7 BLEU points, when using the re-trained Spanish ASR system to transcribe the pSp corpus, instead of only using the baseline ASR system. This result suggests to introduce at least one iteration i in the proposed training scheme, where SLT models are first re-trained with transcriptions from the baseline ASR systems, and then, subsequently trained again with transcriptions from systems that already benefit from re-trained models.

AM	LM_{ASR}	TM		TDev	dev	eval
base	base	base		31.2	25.1	25.3
+92h	+92h	+92h		34.8	28.0	28.3
$+92h_b$	$+92h_b$	$+92h_b$		35.7	28.8	28.4
+92h	+92h	92h		31.8	24.2	24.9
+92h	+92h	$92h_{i=1}$		32.7	25.2	25.6

Table 9. Sp-to-En speech translation results in BLEU.

6. SUMMARY

We explored untranscribed parallel speech audio, as it is present in the multiple audio channel live broadcasts of European Parliamentary Plenary Sessions, as a resource for training SLT systems. Specifically, we showed how SLT training data can be extracted in a fully automatic, unsupervised manner and how this data can then be successfully used to improve the performance of all major models involved in statistical SLT. We applied techniques for unsupervised AM and LM training. Further, we exploited the parallel nature of the given speech audio to train TMs from audio, and to introduce light supervision for SLT model training. We concentrated on a scenario that involves a resource rich language, represented by English, and a resource-deficient language, represented by Spanish. The goal was to improve the statistical models that are affected by the resource-deficiency. Supervised training material for Spanish was limited to 10h of transcribed audio and 100k running words of En/Sp parallel text. With the help of 92h of En/Sp parallel speech audio, we were able to improve the performance of Spanish ASR from 27.1% WER to 23.8%, of Sp-to-En speech translation from 25.3 BLEU to 28.4 and for En-to-Sp speech translation from 21.6 BLEU to 25.2.

Furthermore, we showed that under a strong resource-limitation, where only 10h of transcribed Spanish audio and no parallel text data is available, automatic speech translation and automatic text translation is still feasible with the help of pSp audio. Our Sp-to-En SLT system, based solely on 92h of untranscribed pSp audio (no parallel text) and two TM training iterations, yielded higher translation results than our parallel text trained baseline system.

7. REFERENCES

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

[2] S. Khadivi, A. Zolnay, and H. Ney, "Automatic Text Dictation in Computer-assisted Translation," in *Interspeech*, Portugal, Lisbon, September 2005.

[3] A.Reddy and R. Rose, "Towards Domain Independence in Machine Aided Human Translation," in *Interspeech*, Brisbane, Australia, September 2008.

[4] M. Paulik and A. Waibel, "Extracting Clues from Human Interpreter Speech for Spoken Language Translation," in *Proc. of ICASSP*, Las Vegas, NV, USA, April 2008.

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

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

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

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

[9] L. Lamel, J.L. Gauvain, and G. Adda, "Investigating Lightly Supervised Acoustic Model Training," in *ICASSP*, Salt Lake City, USA, May 2001.

[10] M. Paulik and A. Waibel, "Lightly Supervised Acoustic Model Training on EPPS Recordings," in *Interspeech*, Brisbane, Australia, September 2008.