Automatic Translation from Parallel Speech: Simultaneous Interpretation as MT Training Data

Matthias Paulik and Alex Waibel

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

***Abstract*—State-of-the art statistical machine translation depends heavily on the availability of domain-specific bilingual parallel text. However, acquiring large amounts of bilingual parallel text is costly and, depending on the language pair, sometimes impossible. We propose an alternative to parallel text as machine translation (MT) training data; audio recordings of parallel speech (pSp) as it occurs in any scenario where interpreters are involved. Although interpretation (pSp) differs significantly from translation (parallel text), we achieve surprisingly strong translation results with our pSp-trained MT and speech translation systems. We argue that the presented approach is of special interest for developing speech translation in the context of resource-deficient languages where even monolingual resources are scarce.**

I. Introduction

Translation models (TMs) for statistical MT are traditionally trained from manual translations presented in bilingual, sentence-aligned text corpora. Large amounts of domain-specific, bilingual parallel text are essential to ensure good automatic translation performance. The acquisition of such data can prove time-consuming and costly. This is especially true in the case of resource-deficient languages where the amount of pre-existing bilingual parallel text data is limited. With no or only poorly performing automatic translation available, many bilingual communication scenarios are only possible with the help of interpreters. We propose the use of audio recordings of interpreter speech together with the source language speech, from which the interpretation is being rendered, as an alternative to parallel text as MT training data. Although parallel speech (interpretation) is fundamentally different from parallel text (translation), our experiments show that TMs can be successfully trained from parallel speech alone, without the use of any parallel text data and at surprisingly high performance levels.

Our experiments are conducted on a large-scale simultaneous interpretation task for which we have approximately 100h of pSp available. This enables us to introduce different levels of resource-limitation to examine its effect on MT and speech translation (ST) trained from recordings of pSp. In order to use pSp in a standard training setup for phrase-based statistical MT, we first transcribe both sides of the parallel speech data using automatic speech recognition (ASR). The resulting ASR hypotheses are then aligned on an utterance basis using alignment strategies specifically tailored to the parallel speech of simultaneous interpretation (SI).

We are not aware of any previous work on training TMs from audio recordings of pSp. However, this work is related to and stems from ideas found in previous work on a) improving target language dictation systems for translators from parallel source language text [1], [2], [3], [4], [5] or read speech [3], b) improving speech translation of source language speech for which parallel target language speech is available [6] and c) interpreter speech supervised acoustic model training as used in [7].

II. Translation, Interpretation and PSP

Translation refers to the transfer of meaning from source language text to target language text, with time and access to resources as dictionaries, phrase books, et cetera. Interpretation (as used in the context of this work), refers to the transfer of meaning from source language speech to target language speech. Interpretation happens either simultaneously, while the source language speaker continuously speaks, or consecutively, after the source language speaker has finished speaking. In the latter case, the source and target speech – and the information encoded in that speech – is divided into segments, since interpreter and source speaker wait for their respective turns before speaking. While such a natural segmentation is not given in the context of simultaneous interpretation, the fact that the interpreter has to keep pace with the source language speaker still leads to a time alignment of information in source and target speech.

In the context of this work, we only use SI and no consecutive interpretation. Nevertheless, we define the term parallel speech generally as speech of a source language speaker together with the target language speech of an interpreter. That is, we include simultaneous and consecutive interpretation into this definition. Parallel speech therefore always refers to interpretation. It specifically excludes speech of translators as it was for example used in [3] in the context of dictation systems for translators.

It is important to note that interpretation differs significantly from translation. Interpreters know when and under what circumstances to omit, but also to elaborate and change information and they do not only convey all elements of meaning, but also the intentions and feelings of the source

978-1-4244-5479-2/09/$26.00 © 2009 IEEE

TABLE I
DATA STATISTICS (PARALLEL SPEECH, DEV06, EVAL07)

	English			Spanish		
	pSp	dev	eval	pSp	dev	eval
speech utt.	65.3k	1287	1926	63.2k	1707	2085
transl utt.	65.3k	1194	1167	63.2k	792	746
words [k]	954.4	27.9	26.0	897.0	22.4	25.8
audio [h]	111.3	3.2	2.7	105.2	2.3	2.7

speaker [8]. The differences between SI and translation are strongly influenced by the highly demanding nature of the SI task. In SI, interpreters have to apply special strategies to keep pace with the source language speaker. Corrections of previous interpretation errors, but also fatigue and stress negatively affect the SI quality. Strategies applied during SI include anticipation-strategies [9] and compensatory strategies [10]. For example, interpreters anticipate a final verb or syntactic construction before the source language speaker has uttered the corresponding constituent. The interpreter confirms this anticipation or corrects it when he receives the missing information. The use of open-ended sentences that enable the interpreter to postpone the moment when the verb must be produced is another anticipation-strategy. Compensatory strategies include skipping, approximation, filtering, comprehension omission and substitution. These strategies can lead to a significant loss of information in SI. Experiments reported during the course of the TC-STAR project [11] suggest that the information loss for English-to-Spanish SI as provided during European Parliament Plenary Sessions amounts to approximately 9%. This number was estimated by first creating comprehension questions from an English speech and then determining the number of questions that cannot be answered if only the Spanish SI of the English speech is given. Further, it was reported in [11] that the effective information loss when listening to SI is with 29% significantly higher. This increase in information loss results from the combination of missing information in SI and the difficulty of human evaluators to follow the flow of interpreter speech. In the reported evaluation scenario, the human evaluators were allowed to listen to the recorded interpreter speech twice and they could interrupt the playback to write down their answers.

The strong difference between interpretation and translation can also be expressed in terms of BLEU metric. On our 'dev05' development set (compare Section III-A), we computed a BLEU score of only 14.2, when comparing Spanish-to-English (Sp→En) interpretation with two translation references. For English-to-Spanish (En→Sp), the BLEU score was 18.2. In both cases, we used a manual transcription of the interpreter speech, i.e. the WER was 0%.

III. EXPERIMENTAL SETUP

A. Data and 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 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. In addition to the live broadcasts, the proceedings in the parliament are also published in the form of so-called final text editions (FTEs) in all official languages within approximately two months of the session. These FTEs are created by a multitude of human transcribers and translators from recordings of the original politician's speeches.

For our experiments we use the EPPS part of the English and Spanish TC-STAR spring 2007 verbatim task development and evaluation sets; 'dev06' and 'eval07'. It has to be noted that these sets are only comprised of politician speech and do not include any interpreter speech. This stands in contrast to our pSp corpus. This corpus was collected in-house by recording the European Parliament live broadcast English and Spanish audio channels, which both contain a mix of politician and interpreter speech. ASR acoustic model (AM) training data, as provided during the TC-STAR evaluation campaign, is also comprised of a mix of politician and interpreter speech. Our language models are (mostly) estimated on the FTEs or, in case of the constrained Spanish ASR (compare Section III-B), exclusively on the human reference transcription as used for acoustic model training. Detailed data statistics for dev06, eval07 and the pSp corpus are given in Table I. For dev06 and eval07, the number of speech utterances differs from the number of translation utterances due to the mismatch of automatic speech/non-speech segmentation applied prior to ASR and the manual utterance segmentation of the reference translations. The number of running words in the pSp corpus are estimated on the first best hypotheses of our standard Spanish and English ASR systems presented in Section III-B. The pSp corpus is comprised of sessions from the time periods 04MAY05-26MAY05 and 08SEP05-01JUN06. The development (dev) set has sessions from 06JUN05-06SEP05 and the evaluation (eval) set has sessions from 12JUN06-28SEP06. AM training data is from the time period May 2004 to January 2005. The FTEs are from the time period April 1996 to May 2005.

In addition to the development and evaluation sets, we use one additional Parliamentary session from 26OCT04 to tune our alignment algorithm presented in Section IV-B. We extracted this from the TC-STAR verbatim 2005 dev set. In contrast to the other TC-STAR dev and eval sets used in this work, 'dev05' is a) comprised of a mix of politician and interpreter speech b) forms a pSp corpus and c) is provided with a manual utterance segmentation that is kept consistent for the reference transcriptions and verbatim reference translations. This means in particular, that for all English and Spanish speech utterances, aligned transcription references and translation references are provided. The English (Spanish) side of dev05 consists of 1256 (1589) utterances, with 17.4k (14.7k) running words and 95 (89) minutes of audio.

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

TABLE II
STANDARD ASR SYSTEMS

	dev06		eval07	
	Spanish	English	Spanish	English
PPL	89	108	89	106
WER	8.4	13.9	9.0	12.2

BLEU using two reference translations. We use the multiple reference word error rate (mWER) segmentation script, as it was provided by RWTH Aachen University within TC-STAR, to align the translated speech utterances to the translation references.

B. Automatic Speech Recognition

The employed ASR systems were developed with the Janus Recognition Toolkit (JRTk), featuring the IBIS single pass decoder. The SRI Language Model Toolkit [12] was used for language model (LM) training. Table II gives an overview on the English and the resource-unconstrained Spanish ASR system. A detailed description of these systems is given in the following.

The English ASR system consists of four ASR sub-systems that were developed in our laboratory for the TC-STAR Spring 2006 ASR Evaluation [13]. The decoding setup features a first decoding pass in which two speaker-independent ASR systems with different acoustic front-ends are applied. A traditional Mel-frequency scaled Cepstral Coefficients (MFCC) front-end and a Minimum Variance Distortion-less Response (MVDR) front-end is used. In the same manner, the second decoding pass features two ASR systems with speaker-dependent AMs. Unsupervised speaker adaptation is performed on the output of the previous decoding pass. At the end of both decoding passes, confusion network combination is applied to combine the output of the individual ASR systems. The AM was trained on 80h of English EPPS. The dictionary consists of 47K pronunciation entries. The 4-gram LM was trained on the 2006 available EPPS transcriptions and FTEs, the Hub4 Boradcast News data and the English part of the UN Parallel Text Corpus v1.0.

The decoding setup for the standard Spanish ASR is identical to the setup of the English ASR system. The AMs were trained on 140h of Spanish EPPS and Spanish Parliament (CORTES) data. The dictionary has 74.2K entries. The 4-gram LM was trained on the Spanish FTEs, the CORTES texts and the EPPS + CORTES transcriptions. In addition to the standard Spanish ASR system we use two constrained Spanish ASR systems, cSPASR-0 and cSPASR-1, to simulate ASR performance levels encountered in the context of resource-deficient languages. In the situation of resource-limitation, the lack of text data and transcribed audio data leads to weak LMs and weak AMs. Both contributes to an increased WER. To simulate resource-limitation, we first (cSPASR-0) constrained the Spanish LM to the 748k running words of the transcriptions that were used to train the AM. The constrained LM yields a perplexity (PPL) of 178 on dev06 and a PPL of 177 on eval07, resulting in word error rates of 16.1% and 16.5%, respectively. To simulate a weaker AM (cSPASR-1), we further limited the system to a context independent phone-set. This results in an AM which obtains a WER of 33.3% on dev06 and 33.1% on eval07.

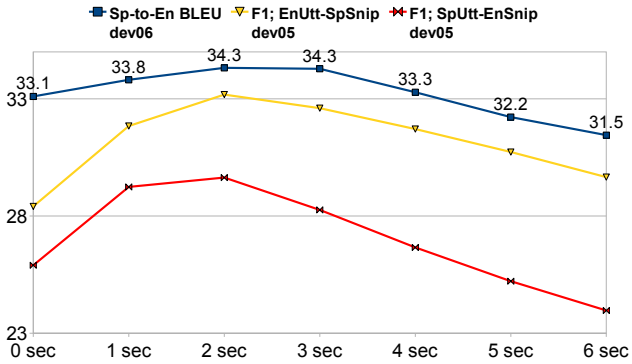

Fig. 1. Dev05 F1-measure and dev06 BLEU for different padding values

C. Machine Translation

For MT, we use the Interactive Systems Labs beam search decoder [14]. The decoder combines multiple model scores to find the best translation. To optimize the system, we use Minimum Error Rate (MER) Training as described in [15].

IV. MT FROM PARALLEL SPEECH

A. Translation Model Training

We use a phrase-based MT approach. Our TMs therefore consist of phrase-to-phrase translation pairs. The phrase tables are trained with the help of the GIZA++ toolkit [16] and University Edinburgh's phrase model training scripts

In contrast to a traditional MT training approach, we do not extract phrase tables from a bilingual, sentence aligned text corpus of manual translations. Instead, we train TMs from interpretation, given in the form of a pSp corpus. In order to do so, we first transcribe both sides of the pSp corpus using ASR and then we introduce an alignment of the resulting ASR hypotheses. This alignment is speech utterance based, meaning that we align to each automatically transcribed source utterance the related target speech ASR transcription. This forms the first part of our translation model training corpus. The second part results from repeating the same alignment procedure in the reverse direction. The alignment procedure is described in detail in the following section.

B. Aligning Interpreter Speech

If not mentioned otherwise, the pSp corpus used in this section is transcribed at an estimated English and Spanish WER level of 12-14% and 9%, respectively. We estimate these numbers based on the English and Spanish ASR performance on dev06 and eval07, since no manual transcription of the pSp corpus is available.

Since simultaneous interpreters have to keep pace with the source language speaker, an implicit time alignment between source and target language speech is already given. We can exploit this fact to align source speech utterances to parallel target speech by considering the target speech snippet that

SPANISH UTTERANCE: "también"
VERBATIM TRANSLATION: "**in addition**"

PARALLEL SPEECH SNIPPET: "and that would mean that we could already start making some of the payments in the year two thousand and five also **in addition** to that we are going to try to make sure that members of staff from different members states of the european union will be granted an equal status"

SPANISH UTTERANCE: "trataremos de que todo el personal tenga"
VERBATIM TRANSLATION: "**we shall try that all the staff will get**"

PARALLEL SPEECH TRANSCRIPT: "in addition to that **we are going to try to make sure that members of staff** from different members states of the european union **will be granted** an equal status we look forward to amend the statuate of course we hope that that will be approved as soon as possible and we hope that it proves viable in practice"

Fig. 2. ± 6 seconds utterance based padding of parallel speech

starts/ends x seconds before/after the source speech utterance starts/ends. We need to include target speech before the start time of the respective source utterance since we do not know which of the audio channels contains interpreter speech. In fact, it often occurs that both audio channels consist of interpreter speech. In such a case the politician that took the floor in the Parliament is giving a speech in a language other than English and Spanish. To minimize computation time, we decode the pSp corpus only once, based on automatic speech utterance segmentation derived via voice activity detection prior to ASR. To extract the ASR hypotheses of the padded speech snippets, we rely on the hypothesized word-start and word-end times.

In order to find an optimal padding value x , we conducted two sets of experiments. First, on dev05 and for different values of x, we computed the F1-measure in respect to uni-gram matches between the padded, automatically transcribed pSp snippets and the for dev05 available reference translations. Figure 1 depicts how the F1-measure changes for different values of x. A peak is obtained at $x = 2$ seconds for both cases, when aligning English utterances to Spanish pSp snippets and when aligning Spanish utterances to English pSp snippets. In the second set of experiments, we created seven different parallel MT training corpora from the automatically transcribed pSp; one training corpus each for $x \in [0-6]$. After extracting seven different phrase tables from these MT training corpora, we computed the translation performance for Sp→En on dev06, using these phrase tables. As we can see in Figure 1, the BLEU score again peaks at $x = 2$, showing that the F1-measure computed on dev05 correlates well with the translation performance on dev06. In other words, the optimal padding value x for aligning our pSp corpus can be well predicted by simply computing the F1-measure on dev05.

In addition to a simple word-time based padding of the parallel speech snippets for aligning the pSp corpus, we also experimented with a more sophisticated two-pass alignment strategy, as presented in the following.

By manually inspecting the pSp present in dev05, we found that, if the information contained in the source utterance is at all present in the pSp, a ± 6 seconds utterance based padding almost always guarantees that the information can be found in the respective target audio snippet. By ± 6 seconds utterance based padding we refer to the case where a target speech snippet is comprised of all target speech utterances that fall into the time window that is formed by padding the source utterance start/end time with 6 seconds. Figure 2 gives an example of pSp that is aligned based on a ± 6 seconds utterance padding. In addition to the transcription reference of the Spanish speech utterance and the respective English pSp-snippet, the Figure shows one of the two Sp→En translation references. The part of the English speech snippet that is directly related to the Spanish speech utterance is marked with an underline. As can be seen, the padded pSp-segment contains too much irrelevant information. The example also shows the strong difference between interpretation and translation.

Our two-pass algorithm for aligning pSp to source speech utterances operates on a per source utterance basis. In its current implementation, the algorithm operates on a first-best ASR hypothesis basis only, but it is scalable to n-best ASR hypotheses. In addition to the source utterance at hand, the algorithm also considers all neighboring source utterances that overlap in their respective target speech snippet with the target speech snippet of the current source utterance. In a first step, the combined forward and backward translation probability for each source word to each target word is computed and an alignment link is introduced if the combined translation probability is above a specific threshold t_p and if the absolute distance between source word start time and target word start time is below a specific threshold t_d. The word-to-word translation probabilities are based on IBM4 word lexicons. The lexicons are trained in a first pass on the parallel MT training corpus that results from a 2 seconds word-time based padding of the pSp snippets. The alignment link between source word sw and target word tw is weighted by the combined translation probability times the 'importance' of the target word. We define the importance of the target word as:

$$importance(tw) = 1.0 - sL * LM(tw) \qquad (1)$$

with sL equal to the length in words of the target speech snippet and $LM(tw)$ equal to the uni-gram LM probability of the target word. In a next step, we find the optimal 'left cut' position in the target speech snippet that defines all words before this position as irrelevant to the source utterance and all words after this position as relevant. This is done by computing the sum over all alignment link weights alw left of a cut position candidate that belong to neighboring source utterances and then adding the sum computed over all alignment link weights alw right of the cut position candidate that belong to the current source utterance. The cut position with the highest overall sum is selected. During this process we also consider alignment link clusters forming target bi- and tri-grams. For each such cluster we introduce additional alignment weights that are included in the overall sum. The alignment weight $alwBI$ for a bi-gram alignment cluster formed by the alignment links $al1$ and $al2$ is for example given as:

$$alwBI(al1, al2) = (alw(al1) * alw(al2))^{bw} \qquad (2)$$

TABLE III
PRECISION, RECALL AND F-MEASURE ON DEV05

n	alignment	EnUtt-SpSnip			SpUtt-EnSnip		
		Pre	Rec	F1	Pre	Rec	F1
1	±2sec	34.8	31.7	33.2	24.9	36.7	29.6
	2-pass	38.9	30.8	34.4	29.7	35.0	32.1
2	±2sec	15.2	12.4	13.7	9.6	13.6	11.2
	2-pass	17.5	12.7	14.7	11.6	13.9	12.7
3	±2sec	8.2	6.4	7.1	4.8	6.8	5.6
	2-pass	9.7	6.6	7.9	5.7	6.9	6.3

TABLE IV
SPANISH-TO-ENGLISH MT PERFORMANCE

	dev06			eval07		
∼WER	9%	16%	33%	9%	16%	33%
±2sec	34.3	32.1	28.2	33.5	32.6	28.5
2-pass	35.1	33.5	29.1	34.3	33.5	30.1

with the bi-gram weight bw to allow for a flexible additional weighting of such bi-gram link clusters. Accordingly an optimal right cut position is found by computing the sum over all alignment links left of the cut position that belong to the current source utterance and adding the sum computed over all alignment links right of the cut position that belong to neighboring source utterances.

To optimize the two-pass alignment algorithm, we performed a grid search on dev05, aiming for a maximal value of F1-measure that is based on matching uni-grams in the pSp snippets and the reference translations. In addition to uni-gram F1-measure (and precision and recall), we also computed the respective values for n-gram matches with $n \in [1-3]$. Table III shows the results for the two alignment passes of the algorithm. The first pass is identical to the 2 seconds word-time based padding of the speech snippets. The table shows that the two-pass algorithm yields higher F1 values at a higher precision and lower recall than the word time based padding. Further, we can see that the overall low recall degrades strongly for higher order n-grams. This underlines the strong difference between translation and interpretation. Table IV lists the Sp→En MT performance when using the two different alignment strategies and automatically transcribing the pSp corpus at different Spanish WER levels. At all three Spanish WER levels, the two-pass alignment strategy outperforms the word-time based alignment by approximately 1 BLEU point. This is in all cases statistically significant ($p < 0.05$). The results also show that, even for a highly degraded Spanish transcription performance at 33% WER (3.7 times worse than the transcription performance of the standard Spanish ASR system), the MT performance degrades only by approximately 12% relative on the eval07 set. This indicates that training TMs from automatically transcribed pSp is robust to strong variations in ASR performance on one side of the pSp corpus.

C. Machine Translation and Speech Translation Results

Table V lists the Sp→En and En→Sp MT results obtained when using TMs trained from pSp. We list the results for pSp that was transcribed at different Spanish WER levels; the approximate Spanish WER achieved on the pSp is shown in the first column. The English ASR system was kept unchanged; we estimate its WER at approximately 12-14% WER, given its performance on dev06 and eval07. BLEU scores marked with $_c$ were computed using the constrained Spanish LM. For comparison, we also list results when training TMs on a bilingual, sentence aligned text corpus of manual translations. This text corpus was extracted from the bilingual MT training corpus as it was provided during the TC-STAR evaluation. We randomly selected sentences pairs from the original TC-STAR training corpus, until the number of running words on the English part reached 954.4k. This is the same number of running words as we estimated for the English part of our pSp corpus. It has to be noted that the TC-STAR training corpus is based on the EPPS FTEs. It therefore exhibits a certain mismatch in style compared to verbatim style transcriptions and translations. To reduce this mismatch we pre-processed the text corpus accordingly.

TABLE V
TRAINING CORPUS DEPENDENT MT PERFORMANCE

training corp.	dev06		eval07	
type, WER	Sp-En	En-Sp	Sp-En	En-Sp
translations, 0%	44.5	41.4	43.9	40.9
interpr., ∼ 9%	35.1	32.8	34.3	31.2
interpr., ∼16%	33.5	27.3_c	33.5	27.0_c
interpr., ∼33%	29.1	24.3_c	30.1	23.3_c

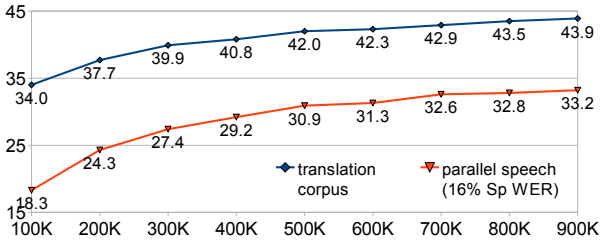

Fig. 3. BLEU score dependent on MT training corpus size

The results show a degraded translation performance for training TMs from pSp, compared to using a bilingual text corpus of manual translations for training. Using our best performing ASR systems, the absolute degradation amounts to approximately 10 BLEU points for both translation directions. This degradation in performance results from a) word errors introduced by automatically transcribing English and Spanish speech b) the mismatch between translation and interpretation and c) errors when aligning the interpreter speech. Nevertheless, we are able to report surprisingly high BLEU scores of up to 34.3 for Sp→En at WER levels of approximately 9% for

TABLE VI
TRAINING CORPUS DEPENDENT ST PERFORMANCE (EVAL07)

training corp.		Sp-En			En-Sp
type, WER		9.0%	16.5%	33.1%	12.2%
translations, 0%		40.0	36.0	27.8	33.8
interpr., ∼ 9%		31.5	-	-	26.1
interpr., ∼16%		-	29.1	-	22.8_c
interpr., ∼33%		-	-	21.0	19.8_c

Spanish ASR and 12-14% for English ASR. As already noted at the end of Section IV-B, we observe only a relatively small degradation in MT performance when introducing a strong degradation in Spanish ASR performance from approximately 9% to 33% WER. In addition to the results listed in Table V, we also computed the MT performance on dev05 for our best pSp-trained models. We achieved BLEU scores of 43.5 and 34.8 for Sp→En and En→Sp, respectively. These scores are 2 to 3 times higher than the BLEU scores we estimated for the dev05 manual transcription of parallel interpreter speech; compare Section II.

The highest achieved Sp→En MT performance on eval07 of 34.3 BLEU is on the same level as the MT performance of TMs trained on 100k English words of sentence aligned translations. We approximate the number of English words in the pSp corpus to be 954.4k (compare Section III-A). Figure 3 depicts the development of BLEU score depending on a successively increased training corpus size in 100k word increments, using either a training corpus of translations or our pSp corpus transcribed at a Spanish WER of 16%. The absolute difference between the BLEU scores of both types of TMs is higher for smaller training corpus sizes. At a corpus size of 100k English words, the difference is 15.7 points (a 46.2% relative degradation) and levels out at 500k words to approximately 10.5 to 11 points (a 24.0% to 26.4% relative degradation).

Table VI lists the speech translation results on eval07. The WER on the respective eval07 source text is shown in the second row. BLEU scores marked with $_c$ were achieved using the constrained Spanish LM. We used the same decoder weights found via MER optimization on the dev06 verbatim transcriptions, as we had good experience in the past with this approach on the very same dev and eval sets. For this reason, we do not provide speech translation results for dev06. Compared to TMs trained on a similarly sized bilingual text corpus of translations, we observe a degradation of approximately 8 BLEU points when using pSp-trained TMs. This degradation in performance is almost 2 BLEU points less than in the case of MT (compare Table V). In general, the relative degradation in BLEU for an increased source input word error rate is smaller for pSp-trained TMs (compare Tables V and VI).

We apply the same ASR systems used for transcribing the pSp corpus when we automatically transcribe the source speech of eval07 for speech translation. The smaller degradation in BLEU score indicates that the pSp-trained TMs are able to compensate source ASR word errors by incorporating mappings between source word errors and their correct target translation. This ability to compensate for source word errors helps to attenuate the loss in speech translation performance experienced by using SI instead of translation for TM training.

V. Conclusion and Future Work

We created a MT training corpus from the untranscribed parallel speech of simultaneous interpreters by automatically transcribing and aligning source language and target language speech. This enabled us to build MT systems and speech translation systems from simultaneous interpretation, thus eliminating the need for a manually created text corpus of sentenced aligned translations. We achieve surprisingly strong translation results with our pSp-trained translation models of up to 34.3 (31.2) BLEU points for Sp→En (En→Sp) MT and up to 31.5 (26.1) BLEU points for Sp→En (En→Sp) speech translation. Our experiments show that training TMs from pSp is robust towards low transcription performance on one side of the automatically transcribed speech corpus. Therefore, we argue that simultaneous interpreter speech can present a valuable resource for training MT and speech translation in the context of resource-deficient languages. Furthermore, our experiments show that in the case of speech translation, pSp-trained TMs profit from an ability to compensate for word errors in the source ASR.

We will continue to explore interpretation as a resource for training MT and speech translation. We plan to expand our experiments to consecutive interpretation and to additional language pairs.

References

[1] P. Brown, S. Chen, S. DellaPietra, V. DellaPietra, A. Kehler, and R. Mercer, "Automatic Speech Recognition in Machine Aided Translation," *Computer Speech and Language*, vol. 8, no. 3, pp. 177–87, 1994.

[2] J. Brousseaux, G. Foster, P. Isabelle, R. Kuhn, Y. Normandin, and P. Plamondon, "French Speech Recognition in an Automatic Dication System for Translators: the TransTalk Project," in *Eurospeech*, Madrid, Spain, September 1995.

[3] M. Paulik, S. Stüker, C. Fügen, T. Schultz, T. Schaaf, and A. Waibel, "Speech Translation Enhanced Automatic Speech Recognition," in *Proc. of ASRU*, San Juan, Puerto Rico, 2005.

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

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

[6] 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.

[7] ——, "Lightly Supervised Acoustic Model Training on EPPS Recordings," in *Interspeech*, Brisbane, Australia, September 2008.

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

[9] F. V. Besien, "Anticipation in Simultaneous Interpretation," *Meta: Journal des traducteurs*, vol. 44(2), pp. 250–259, 1999.

[10] R. Al-Kahnjii, S. El-shiyab, and R. Hussein, "On the Use of Compensatory Strategies in Simultaneous Interpretation," *Meta: Journal des traducteurs*, vol. 45(3), pp. 544–557, 2000.

[11] D. Mostefa, O. Hamon, N. Moreau, and K. Choukri, "TC-STAR Evaluation Report, Del.no 30," www.tc-star.org, May 2007.

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

[13] S. Stüker, C. Fügen, R. Hsiao, S. Ikbal, Q. Jin, F. Kraft, M. Paulik, M. Raab, Y. Tam, and M. Wölfel, "The ISL TC-STAR Spring 2006 ASR Evalutation Systems," in *TC-STAR Workshop*, Barcelona, Spain, 2006.

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

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

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