EXTRACTING CLUES FROM HUMAN INTERPRETER SPEECH FOR SPOKEN LANGUAGE TRANSLATION

Matthias Paulik and Alex Waibel

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

ABSTRACT

In previous work, we reported dramatic improvements in automatic speech recognition (ASR) and spoken language translation (SLT) gained by applying information extracted from spoken human interpretations. These interpretations were artificially created by collecting read sentences from a clean parallel text corpus. Real human interpretations are significantly different. They suffer from frequent synopses, omissions and self-corrections. Expressing these differences in BLEU score by evaluating human interpretations with carefully created human translations, we found that human interpretations perform two to three times worse than state-of-the art SLT. Facing these stark differences, we address the question if and how ASR and SLT can profit from human interpretations. In the following we describe initial experiments that apply knowledge derived from real human interpretations for improving English and Spanish ASR and SLT. Our experiments are conducted on a small European Parliamentary Plenary Sessions development set.

Index Terms— spoken language translation, STE-ASR, tight coupling

1. INTRODUCTION

As of today, European Parliamentary Plenary Sessions (EPPS) are broadcast live in up to 22 languages. The respective original-language texts, along with their provisional translations, are published after several weeks. These provisional translations are subsequently replaced by their so-called final text editions. This tedious and expensive process may be supported effectively by transcriptions and translations that are automatically created from the broadcast audio. In this work, we examine the possibility of improving automatic speech recognition (ASR) and spoken language translation (SLT) applied to EPPS by extracting information from the available human interpretations. We report promising results for initial experiments conducted on a small English↔Spanish EPPS development set.

In previous work [1], we showed dramatic improvements in ASR and SLT when incorporating information from human interpretations. These interpretations were artificially created by collecting sentences read from the bilingual Basic Travel Expression Corpus (BTEC). In our current work we use for the first time real human interpretations, as they are provided during EPPS. These real human interpretations suffer from frequent synopses, omissions and self-corrections. Further, the assumption made in our previous work, that every spoken source sentence comes with a perfectly aligned spoken target sentence is no longer valid. This assumption enabled us to directly tailor ASR individually for each sentence. This sentence based tailoring was done mostly by rescoring ASR n-best lists with translation model scores and by favoring words found in the automatic translations through a manipulated ASR language model (LM). The improved English and Spanish recognition hypotheses were also used to adapt the involved machine translation (MT) systems towards the transcriptions of the respective target sentences. With these adapted MT systems, it was possible to further improve ASR performance. This lead to an overall iterative system of mutual ASR and MT adaptation as depicted in the left hand side of Figure 1. We called this approach Speech Translation Enhanced Automatic Speech Recognition (STE-ASR).

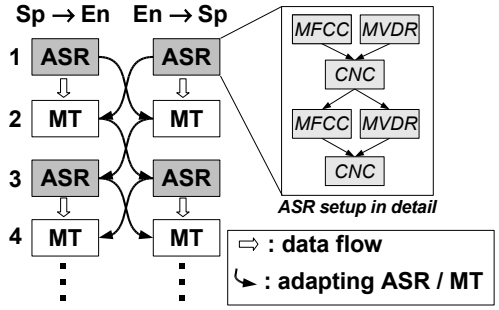

Fig. 1. Planned overall system architecture.

2. EXPERIMENTAL SETUP

2.1. Planned Overall System Architecture

The overall system architecture for our experiments is depicted in Figure 1. This architecture allows for a mutual adaptation of all involved SLT components. It is based on the system architecture described in [1]. In step 1, we automatically transcribe the English and Spanish speech. Step 2 consists of using the ASR hypotheses to bias the En↔Sp MT systems and then to automatically translate the ASR hypotheses using the biased MT systems. The information gained from the translated ASR hypotheses is used in a third step to bias the ASR systems and to produce new, improved transcriptions. These steps can now be repeated until no further improvement in performance is observed. In this work, we report initial experiments

This work has been funded in part by the European Union under the integrated project TC-Star -Technology and Corpora for Speech to Speech Translation - (IST-2002-FP6-506738, http://www.tc-star.org). This work is also partly supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-2-0001. 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.

conducted for step 1 to 3.

2.2. Data

Our experiments were conducted on one of the European Parliament Plenary Sessions included in the official TC-STAR [2] 2005 development set. The segmented English audio track of the session had 2236 utterances with 108 minutes of audio. The Spanish track consisted of 1728 utterances with 98 minutes of audio. Speakers in both audio tracks change between politicians and human interpreters. In the case that a politician is speaking neither in English nor in Spanish, the speakers in both tracks are human interpreters.
SLT performance is measured in BLEU computed towards two case sensitive translation references without punctuation. It is important to notice that the SLT references are not identical to the transcription references of the respective parallel speech.

2.3. ASR Baseline Systems

The employed ASR systems were developed with the help of the Janus Recognition Toolkit (JRTk), featuring the IBIS single pass decoder. They consist of sub-phonetically tied three-state Hidden Markov Models (HMMs) without state skipping. Acoustic model training involved incremental splitting of Gaussians followed by the estimation of one global semi-tied covariance matrix after LDA and several iterations of Viterbi training. The SRI Language Model Toolkit was used for language model (LM) training.

The overall **English ASR** system consists of ASR sub-systems taken from the ISL TC-STAR Spring 2006 ASR Evaluation systems [3]. These ASR sub-systems are tied together in a decoding setup as depicted on the right hand side of Figure 1. The 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) [4] front-end is used. The second decoding pass features two ASR systems with speaker-dependent acoustic models. Unsupervised speaker adaptation is performed on the output of the first decoding pass. At the end of both decoding passes, confusion network combination (CNC)[5] is applied to combine the output of the individual ASR systems. The acoustic models were trained on 80h of English EPPS. The pronunciation dictionary consists of 47K lowercased pronunciation entries. The 4-gram language model (LM) was trained on the 2006 available EPPS transcriptions and EPPS final text editions, the Hub4 Boradcast News data and the English part of the UN Parallel Text Corpus v1.0. The LM perplexity on the development session is 80 and the (case-sensitive) WER is 13.1% (15.2%). Capitalized recognition output, as it is used as input to our En→Sp MT system, is created in a post-processing step with the help of a case-sensitive 4-gram LM trained on the EPPS corpus.

We developed the **Spanish ASR** system using the same techniques as were used to develop the English ASR system. The overall decoding setup is the same, i.e. it is identical to the setup shown in Figure 1. The acoustic models were trained on 140h of Spanish EPPS and Spanish Parliament (CORTES) data. The pronunciation dictionary has 77.9K entries over a case-sensitive vocabulary of 63.3K. The case-sensitive 4-gram LM was trained on the Spanish EPPS final text editions, the CORTES texts and the EPPS + CORTES transcriptions. The LM perplexity on the development session is 71 and (case-sensitive) WER is 10.7% (11.8%).

Sp→En (csWER 0%)	Sp→En (csWER 11.8%)	En Transcript
54.69	48.77	14.61
En→Sp (csWER 0%)	**En→Sp (csWER 15.2%)**	**Sp Transcript**
45.92	40.94	19.8

Table 1. BLEU scores of the baseline SLT and the transcription references. The case-sensitive WER of the source is shown in brackets.

2.4. MT Baseline Systems

The used En↔Sp MT systems are based on MT systems developed in our laboratory for the TC-STAR Spring 2007 SLT Evaluation. The systems were trained on the parallel EPPS corpus. Phrase tables were estimated via the GIZA++ toolkit [6] and University of Edinburgh's phrase model training scripts. The phrase tables consist of phrase-to-phrase translations annotated with four TM scores. These four scores are the forward and backward phrase translation probabilities and the forward and backward lexical weights. Both systems apply a 4-gram LM built with the SRI LM toolkit and a 6-gram suffix array LM. The LMs were trained on the respective target side of the EPPS corpus. The ISL beam search decoder [7] combines the TM and LM scores together with scores from an internal word reordering model and simple word and phrase count models to find the best translation. To optimize the system towards a maximal BLEU score, we use minimum error rate (MER) training as described in [8]. For each model weight, MER applies a multi-linear search on the development set n-best list produced by the system. The estimated model weights are used to produce new translation n-best lists and the process is repeated until the translation quality converges.

3. CAN WE GAIN FROM HUMAN INTERPRETATIONS?

The following shows the Sp→En SLT references along with the English transcription reference for the beginning of the used EPPS:

Sp → En ref. 1: 'I would like to ask if there are any remarks on the minutes from yesterday's sitting that I trust are in your possession no remarks all right we will then consider the minutes to be approved'

Sp → En ref. 2: 'I would like to ask you whether you have any observation to make with respect to the records of the yesterday session which I hope are in your hands no observation fine then we shall consider those records approved'

En transcript: 'ladies and gentlemen the sitting is open you have received the minutes from yesterday's meeting I wanted to ask you whether you have any comments on the minutes of yesterday's meeting if not and there do not seem to be any comments in that case the minutes can be deemed approved'

When comparing these references it becomes immediately apparent that they differ significantly. Tables 1 compares the BLEU scores of our baseline SLT systems to the BLEU scores of the English and Spanish ASR transcription references for the complete session. While the transcriptions have to be considered valid (simultaneous) translations, their translation performance in BLEU measured towards the two SLT references is two to three times worse than the performance of the SLT systems. This huge difference can

be explained by certain strategies applied by the human interpreter to keep up with the source language speaker. These strategies include frequent synopses and planned omissions. More frequently occurring self-corrections in the interpreter's speech as well as accidental omissions may also play a major role.
Regarding this 'poor' translation performance it is questionable how 'useful' information extracted from human interpreter speech is in the context of improving ASR and SLT performance.

4. IMPROVING SLT PERFORMANCE

In the framework of the planned overall system as it is described in Figure 1, an improved SLT can be interpreted in two ways. With the goal of having a best possible ASR performance in mind, it can be argued that an improved SLT performance is gained whenever the translation is improved in the sense of providing best possible information for a further successful ASR adaptation. In this context, an improved SLT performance should be reached by biasing the system as strong as possible towards the transcript of the respective parallel audio, without incorporating ASR errors into the MT systems. We will refer to such a kind of improvement as a 'relative' SLT improvement. Accordingly, we call an improvement of the SLT systems towards the provided SLT references an 'absolute' SLT improvement.

4.1. Extracting ASR n-gram hints

We investigated two different strategies, as well as their combination, to incorporate information from the ASR hypotheses into the MT systems. First, we extracted n-grams with $n = 1, 2, 3$ from the ASR first-best hypotheses of the target audio snippet that starts/ends 6 seconds before/after the start/end time for each source language utterance. We call these n-grams ASR n-gram hints. These hints are loaded during run-time by our translation component before translating a given source utterance. Whenever an ASR n-gram hint is observed during decoding, a discount is applied to the score (cost) of the current translation hypothesis. In this way, we favor translations that contain ASR n-gram hints. All ASR n-grams of the same order share the same discount value. The specific discount value is estimated via MER optimization. MER optimization was performed towards the two SLT references, using the transcription references as well as the ASR first-best hypotheses as input. Due to alignment problems and decoder constraints, we did not perform a MER optimization towards the transcriptions of the parallel audio.

4.2. Extracting ASR phrases

The second approach we investigated extracts complete translation phrases from the source and target language ASR hypotheses and incorporates them, again dynamically for each source utterance, into the baseline phrase tables of the MT systems. We extract ASR translation phrases by computing the alignment matrix between source ASR first-best hypothesis and target ASR first-best hypothesis of the 6 seconds padded target audio snippet. In a first iteration, this alignment matrix consists only of word-to-word translation probabilities extracted from the forward and backward IBM4 lexicons of the MT systems. We then estimate for each source word a discrete probability distribution for source-to-target word delays d, with $d \in [-6, .., 0, .. + 6]$ seconds. The source-to-target word delay is defined as the distance in seconds between the start time of the source words and their respective target language translation in the parallel audio. For estimating the discrete probability distribution, we consider only words that are aligned with a high lexical translation probability and that are found within a 60 second window around the current source word. The alignment matrix is then re-estimated, using an interpolation of lexical translation probabilities and the estimated delay alignment probability. In a next step, we introduce binary alignment links. These binary alignment links are computed with the help of a simple algorithm described in [9]. This algorithm allows limited alignment link overlaps, i.e. links that either share the same source or the same target word. In a final step, we cluster the binary alignment links using a neighborhood of k source and target words around each link. These clusters now constitute ASR translation phrases. One example for a phrase extracted in this manner, with a neighborhood of $k = 1$, is: *presupuesto aqui en el Parlamento # budget here in the Parliament*. For each of these phrases, we compute the forward and backward translation probability based on the IBM4 lexicons. In order to be able to incorporate these new phrases into the baseline phrase table, we extend the baseline phrase entries with two additional TM probabilities that are set to 1 (zero logarithmic cost). Accordingly, the additional ASR phrases have probabilities of 1 at the positions of the four original TM probabilities, followed by the two computed TM probabilities of the ASR phrases. The optimal weights by which these different translation model probabilities contribute to the translation score (cost) used by the decoder are estimated via MER optimization.

Translation Direction	En → Sp		Sp → En	
csWER of source	0.0%	15.2%	0.0%	11.8%
Baseline	45.92	40.94	54.69	48.77
ASR n-grams	46.48	41.01	55.06	49.48
ASR phrases	46.30	41.05	54.71	48.76
ASR n-gram & phrases	46.52	40.91	54.99	49.58

Table 2. 'Absolute' SLT improvement measured in BLEU towards the two official SLT references.

Translation Direction	En → Sp		Sp → En	
csWER of source	0.0%	15.2%	0.0%	11.8%
Baseline	12.84	12.10	8.49	7.88
ASR n-grams	13.89	12.81	9.94	9.29
ASR phrases	13.07	12.86	8.50	7.80
ASR n-gram & phrases	14.34	14.85	9.85	9.85

Table 3. 'Relative' SLT improvement measured in BLEU towards the transcription references of the parallel audio.

4.3. Experimental Results

Table 2 shows the 'absolute' SLT translation performance in BLEU score when incorporating ASR n-gram hints, ASR phrases and a combination of both. While we see small but consistent improvements for incorporating ASR n-gram hints, applying ASR phrases only gives a noticeable improvement for En→Sp when using the English transcription references as input. In regards to an additional improvement when adding ASR phrases on top of ASR n-gram hints, we observe the most noticeable improvements in both translation directions when translating the human transcription references. The ASR phrases are extracted from the first-best ASR hypotheses, i.e. the phrases contain recognition errors found in these hypotheses. Translating the transcription references prevents ASR phrases with recognition errors on the source side (and therefore potentially wrongly aligned phrases) from being selected, which explains the better performance for this case.
While the improvements in terms of an 'absolute' SLT performance measured against the two SLT references are small, it is arguable

	Cheating Filtering	Agreement Filtering	
	Baseline MT	Baseline MT	Improved MT
1-gram	100 - 54 - **70**	90 - 48 - **62**	90 - 50 - **65**
2-gram	100 - 23 - **37**	88 - 19 - **31**	88 - 24 - **37**
3-gram	100 - 11 - **20**	85 - 9 - **16**	86 - 13 - **23**

Table 4. Precision, Recall and **F-Measure** of the filtered MT n-gram hints for English ASR improvement.

	MFCC	MVDR	**CNC**
Baseline	13.7%	13.5%	**13.1%**
Cheating	12.4%	11.9%	**11.6%**
Agreement	13.5%	13.1%	**12.8%**

Table 5. English ASR improvements in lowercase WER.

if the translation performance in itself may have improved more strongly in the sense of a valid translation that is closer to the parallel audio transcription. A first indication for such a 'relative' improvement can be found when computing the BLEU score of the MT systems towards the human transcript of the parallel audio. Table 3 lists the according BLEU scores. Although we did not perform a MER optimization towards these references, we see consistent gains for applying a combination of ASR n-gram hints and ASR phrases. However, in the context of this work, we achieve this 'relative' SLT performance mostly having the goal of a further improvement of the involved ASR systems in mind. For this reason, we will measure the 'relative' SLT performance in terms of an absolute improvement of ASR performance as well as in terms of F-measure regarding the quality of the extracted MT n-gram hints used for ASR improvement, as described in the next section.

5. IMPROVING ASR PERFORMANCE

To bias ASR, we apply MT n-gram hints with $n = 1, 2, 3$. MT n-gram hints are n-grams found in the m-best translations $(m = 500)$ of the target ASR first-best hypothesis. The ASR hypothesis is computed from the 6 seconds padded target audio snippet. In order to apply these hints to ASR, we changed our decoder to be able to manipulate n-gram LM probabilities dynamically during run-time. Our ASR decoder now loads 3 different 'discount' language models of order 1, 2 and 3, in addition to the baseline ASR LM. Initially, these discount LMs have a logarithmic probability (cost) of 0 for all n-grams. Before decoding an utterance, we change the cost of the MT n-grams for the current source utterance to their respective discount value. During decoding time, we subtract the scores provided by these discount LMs from the score of the baseline ASR LM. This mechanism enables us to apply MT n-gram discounts during decoding time to the score of the current (partial) ASR hypothesis. MT n-gram hints of the same order share the same discount value, which is estimated via a manual grid search.

5.1. Experimental Results

For a first evaluation of the usefulness of the MT n-gram hints for ASR improvement, we performed a cheating experiment with the English ASR system. In this experiment, we filtered the available MT n-gram hints with the human source transcriptions, i.e. we kept only those hints that actually occur in the reference transcription. One way to measure the quality of the resulting MT n-gram hints without having to compute WER is to compute the F-measure of these hints. The row labeled 'cheating' in Table 4 lists the precision, recall and F-measure of the filtered Sp→En MT n-gram hints. Since we filtered the hints with the transcription reference itself, we have a precision of 100%. Applying these cheating hints during the CNC step of the second pass of the English ASR reduces the case-insensitive WER from 13.1% to 11.7%. Recreating the lattices used as input to the CNC step, with the cheating hints applied during lattice creation, leads to an additional drop of 0.1% in WER for the final result after CNC. The column labeled 'cheating' in Table 5 lists detailed results for the latter case.

Transferring the concept of filtering the MT n-gram hints to a realistic experiment leads to a filtering based on ASR n-best hypotheses. We label this concept 'agreement' filtering of the MT n-gram hints. We only keep MT n-grams found in the MT 500-best and the ASR 250-best hypotheses. Tables 4 and 5 list the detailed results in F-measure and WER for these agreement hints.

6. CONCLUSION

We examined the English and Spanish parallel audio streams provided during live broadcasts of European Parliamentary Plenary Sessions for their usability in improving ASR and SLT applied to the very same audio. Our initial experiments showed that it is possible to extract information from the parallel audio streams to benefit such ASR and SLT systems. In particular, we were able to reduce the English ASR WER by 0.3% absolute from 13.1% to 12.8%. SLT improved by 0.8 BLEU points for Sp→En and by 0.1 BLEU points for En→Sp. The baseline BLEU scores were 48.77 and 40.94, respectively. When measuring SLT improvement using the human transcription of the parallel audio as reference, the gains were 2.8 BLEU points for En→Sp and 2 BLEU points for Sp→En. The baseline BLEU scores were 12.10 and 7.88, respectively.

7. REFERENCES

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

[2] TC-STAR, "Technology and Corpora for Speech to Speech Translation," http://www.tc-star.org.

[3] 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 *Proc. of the TC-STAR Workshop on Speech-to-Speech Translation*, Barcelona, Spain, 2006.

[4] M.C. Wölfel and J.W. McDonough, "Minimum Variance Distortionless Response Spectralestimation, review and refinements," in *IEEE Signal Processing Magazine*, Antwerp, Belgium, 2005, pp. 117–126.

[5] L. Mangu, E. Brill, and A. Stolcke, "Finding Consensus in Speech Recognition: Word Error Minimization and Other Applications of Confusion Networks," in *Computer Speech and Language*, April 2000, pp. 373–400.

[6] F.J. Och and H. Ney, "Improved Statistical Alignment Models," in *Proc. of ACL*, Hongkong, China, 2000.

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

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

[9] Jörg Tiedemann, "Combining Clues for Word Alignment," in *Proc. of ACL*, Budapest, Hungary, 2003.