Document Driven Machine Translation Enhanced ASR

M. Paulik[1,2]*, C. Fügen*[1]*, S. Stüker*[1]*, T. Schultz*[2]*, T. Schaaf*[2]*, and A. Waibel*[1,2]

[1]Institut für Logik, Komplexität und Deduktionssysteme, Universität Karlsruhe (TH)
[2]Interactive Systems Laboratories, Carnegie Mellon University, Pittsburgh
{paulik|fuegen|stueker|waibel}@ira.uka.de, {tschaaf|tanja}@cs.cmu.edu

Abstract

In human-mediated translation scenarios a human interpreter translates between a source and a target language using either a spoken or a written representation of the source language. In this paper we improve the recognition performance on the speech of the human translator spoken in the target language by taking advantage of the source language representations. We use machine translation techniques to translate between the source and target language resources and then bias the target language speech recognizer towards the gained knowledge, hence the name Machine Translation Enhanced Automatic Speech Recognition. We investigate several different techniques among which are restricting the search vocabulary, selecting hypotheses from n-best lists, applying cache and interpolation schemes to language modeling, and combining the most successful techniques into our final, iterative system. Overall we outperform the baseline system by a relative word error rate reduction of 37.6%.

1. Introduction

Human-mediated translation refers to situations in which a speaker of one language communicates with one or several speakers of another language with the help of a bilingual human interpreter who mediates between the communication partners. One example is an American aid worker who speaks with a non-American victim through a human interpreter. Another example is a Spanish speaker delivering a speech to a non-Spanish audience. In the latter example one (or several) interpreters would translate the Spanish spoken presentation into the language(s) of the listeners. This happens either directly from the spoken speech or with the help of a transcript of the delivered speech. In both examples it is desirable to have a written transcript of what was said by the interpreter, e.g. for archiving and retrieval, or publication. The most straight-forward technique is to record the speech of the interpreter and then use automatic speech recognition (ASR) to transcribe the recordings. Since additional knowledge in form of a spoken and/or a written representation of the source language is available it can be used to improve the performance of the ASR. One possibility is the use of machine translation (MT) to translate these resources from the source into the target language as illustrated in Figure 1.

Dymetman et al. [1] and Brown et al.[3] proposed this approach in 1994. In the TransTalk project [1, 2] Dymetman and his colleagues improved the ASR performance by rescoring the ASR n-best lists with a translation model. Furthermore, they used the translation model to dynamically create a sentence-based vocabulary list in order to restrict the ASR search space. In [3] Brown et al. introduce a technique for applying the translation model during decoding by combining its probabilities with those of the language model. Applying a similar idea as [1], Placeway and Lafferty [4] improved the recognition accuracy on TV broadcast transcriptions using closed-captions. Ludovik and Zacharski show in [5] that using MT for constraining the recognition vocabulary is not helpful but that good improvements can be observed by using a MT system for topic detection and then choosing an appropriate topic specific language model for recognition.

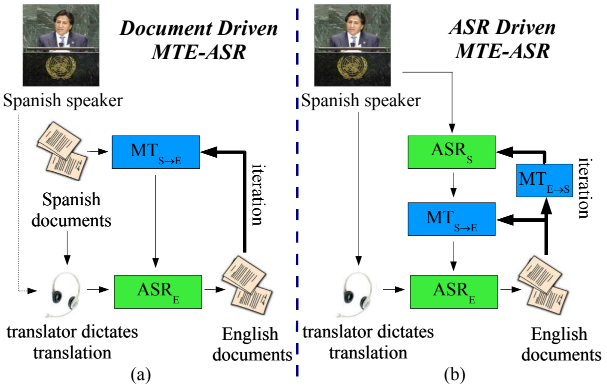

Figure 1: MT-Enhanced ASR

Our work goes beyond the described research by developing an iterative system that incorporates all knowledge sources available for both - the source and target language, and by optimizing the integrated system. Figure 1 (a) depicts the overall iterative system design in the case of available source documents. The following experiments refer to this case only. The key idea of this system design is that, in the same manner as it is possible to improve the ASR system in the target language, it is possible to improve the performance of the source MT system, for example by using the translation transcription together with the source documents as additional training data. This motivates the shown feedback loop.

2. System Component Improvements

In this chapter we compare different techniques to improve the performance of the system's main components. In particular we describe techniques to improve the ASR component using knowledge provided by the MT component, and techniques to improve the MT component using knowledge derived from ASR. The performance improvements on the ASR are described in terms of word error rates (WERs) and were gained by using the baseline MT knowledge only, i.e. without iterations. While for the experiments on the MT component we used the improved ASR output corresponding to the first iteration of the document driven MTE-ASR system.

2.1. Data Set for Component Evaluation

For the evaluation of the two system components, ASR and MT, we used a data set consisting of 506 parallel Spanish and English sentences taken from the bilingual Basic Travel Expression Corpus (BTEC). The 506 English sentences were presented four times, each time read by different speakers. After removing some corrupted audio recordings, a total of 2008 spoken utterances (798 words vocabulary size) or 67 minutes speech from 12 different speakers were derived as the final data set. Since each sentence is spoken four times by four different speakers, we split the ASR output into disjoint subsets, such that no subset has the hypothesis /n-best list of the same sentence spoken by different speakers. Based on these four subsets we trained four different MT components. The presented performance numbers reflect the average performance calculated over the four results.

2.2. Baseline Components

2.2.1. English ASR

For the ASR experiments in this work we used the Janus Recognition Toolkit (JRTk) featuring the IBIS single pass decoder [6]. Our sub-phonetically tied three-state HMM based recognition system has 6000 codebooks, 24000 distributions and a 42-dimensional feature space on MFCCs after LDA. It uses semi-tied covariance matrices, utterance-based CMS and incremental VTLN with feature-space MLLR. The recognizer was trained on 180h Broadcast News data and 96h Meeting data [7]. The back off tri-gram language model was trained on the English BTEC which consists of 162.2 K sentences with 963.5 K running words from 13.7 K distinct words. The language model perplexity on the data set described above was 21.6. The OOV rate was 0.52%. The system parameters were tuned on the complete data set. The word error rate (WER) was 12.6%.

2.2.2. Spanish to English MT

The ISL statistical machine translation system [8] was used for the Spanish to English automatic translations. This MT system is based on phrase-to-phrase translations (calculated on word-to-word translation probabilities) extracted from a bilingual corpus, in our case the Spanish/English BTEC. It produces a n-best list of translation hypotheses for a given source sentence with the help of its translation model (TM), target language model and translation memory. The translation memory searches for each source sentence that has to be translated the closest matching source sentence, with regard to the edit distance, in the training corpus and extracts it along with its translation. In case of an exact match the extracted translation is used, otherwise different repair strategies are used to find the correct translation. The TM model computes the phrase translation probability based on word translation probabilities found its statistical IBM1 forward and backward lexica regardless of the word order. The word order of MT hypotheses is therefore appointed by the LM model and translation memory. As the same LM model is used as in the ASR baseline system one can say that only the translation memory can provide additional word order information for ASR improvement. The system gave a NIST score of 7.13, a BLEU score of 40.4.

2.3. Experiments and Results on ASR

2.3.1. Vocabulary Restriction

In our first experiment we restricted the vocabulary of the ASR system to the words found in the MT n-best lists. For an MT n-best list of size n=1 a WER of 26.0% was achieved, which continuously decreased with larger n, reaching 19.6% for n=150. We computed a lower bound of 15.0% for $n \rightarrow \infty$ by adding all OOV words to the n=150 vocabulary. This means that no improvement in recognition accuracy could be achieved by this vocabulary restriction approach.

2.3.2. Hypothesis Selection by Rescoring

The n-best WER (nWER) found within the ASR 150-best lists of the baseline system is 6.5%. This shows the huge potential of rescoring the ASR n-best lists. In contrast, the best WER that can be achieved on the 150-best MT list is 34.2%. However, when combining the n-best lists of ASR and MT the nWER reduced to 4.2% which proves that complementary information is given in the n-best lists of both components. In fact, we observed the best rescoring performance when enriching the ASR 150-best list with just the first best MT hypothesis. Therefore, all mentioned rescoring results refer to in this manner enriched ASR n-best lists. The applied rescoring algorithm computes new scores (negative log-probabilities) for each sentence by summing over the weighted and normalized translation model (TM) score, language model (LM) score, and ASR score of this sentence. To compensate for the different ranges of the values for the TM, LM and ASR scores, the individual scores in the n-best lists were scaled to $[0;1]$.

$$s_{final} = s'_{ASR} + w_{TM} * s_{TM} + w_{LM} * s_{LM} \qquad (1)$$

The ASR score output by the JRTk is an additive mix of acoustic score, weighted language model score, word penalty and filler word penalty. The language model score within this additive mix contains discounts for special words or word classes. The rescoring algorithm allows to directly change the word penalty and the filler word penalty added to the acoustic score. Moreover, four new word context classes with their specific LM discounts are introduced: MT mono-, bi-, trigrams and complete MT sentences. MT n-grams are n-grams included in the respective MT n-best list; MT sentences are defined in the same manner. The ASR score in equation (1) is therefore computed as:

$$\begin{aligned} s'_{ASR} = & s_{ASR} + lp' * n_{words} + fp' * n_{fillerwords} \\ & - md * n_{MTmonograms} - bd * n_{MTbigrams} \qquad (2) \\ & - td * n_{MTtrigrams} - sd * \delta_{isMTsentence} \end{aligned}$$

Parameter optimization was done by manual gradient descent. The best parameters turned out to be w_{TM}=0.2, w_{LM}=0.4, md=58, fp'=-35, and all other parameters are set to zero. This system yielded a WER of 10.5% which corresponds to a relative gain of 16.7%. The MT is not able to produce/score nonlexical events seen in spontaneous speech. This accounts for the negative rescoring filler penalty of fp'=-35: the ASR score has to compete with the filler penalty free TM (and LM) score during rescoring. This approach offers a successful way to apply MT knowledge for ASR improvement without changing the ASR system. MT knowledge is applied in two different ways: by computing the TM score for each individual hypothesis and by introducing new word class discounts based on MT n-best lists. The fact that of the word class discount parameters only the mono-gram discount is different from zero, shows that the word context information provided by the MT is of little value for the ASR. On the other hand, the mono-gram discount contributes largely to the success of this approach: the best WER found without any discounts was 11.50%. Thus the MT is not very useful to get additional word context information, but very useful as a provider for a ”bag of words”, that predicts which words are going to be said by the human translator.

2.3.3. Cache Language Model

Since the mono-gram discounts have such a great impact on the success of the rescoring approach it is desirable to use this form of MT knowledge not only after, but already during ASR decoding. This will influence the pruning applied during decoding in a way that new, correct hypotheses are found. In our cache LM approach we define the members of the word class mono-gram in the same manner as above, but now dynamically, during decoding. The best performing system uses MT n-best lists of size n=20 and a log probability discount of d=1.3. This procedure yielded a WER of 10.4% and had therefore a similar performance as the rescoring approach. But in contrast to the rescoring approach only two parameters are used. Moreover, the expectation to find new, correct hypotheses could be fulfilled: the nWER for the Cache LM system output was now 5.5% in comparison to 6.5% of the baseline system.

2.3.4. Language Model Interpolation

In this experiment the language model of the baseline ASR system was interpolated with a small language model computed on the translations found in the MT n-best lists. The best system has an interpolation weight of i=0.2 for the small MT language model and a MT n-best list size of n=30. The resulting WER was 11.6%. When using a sentence based interpolation instead, i.e for each sentence a small LM is computed on the respective MT n-best list, the WER increased to 13.2%. The LM interpolation approach uses MT context information in form of tri-grams (and bi- and mono-grams for backoff). The, in comparison to the rescoring and cache LM approach, small gain in WER can be explained by the already stated little value of MT context information.

2.3.5. Combination of ASR Improvement Techniques

The introduced ASR improvement techniques apply different forms of MT knowledge with varying success. Therefore, we examined if it is possible to further increase the recognition accuracy by combining these techniques:

Cache LM on Interpolated LM: Combining the cache and interpolated LM schemes a minimal WER of 10.1% was obtained for the cache LM parameters n=20, d=1.4 and interpolation LM parameters i=0.1, n=60. This is only a small improvement compared to the cache LM. Once again we can argue that the MT context information used within the interpolated LM is of little value and that the success of the interpolated LM approach is largely due to the mono-gram backing-off. As the cache LM approach is already based on MT knowledge provided through MT mono-grams the combination with the interpolated LM can only yield small improvements.

Hypothesis Selection on Cache LM System Output: For this experiment the above described rescoring algorithm was used on the n-best lists produced by the best found cache LM system. The best WER found was 9.4% when using the parameter setting w_{TM}=0.075, w_{LM}=0.025, bd=2, sd=2, fp'=-20, lp'=5, n_{ASR}=150, n_{MT}=1 and all other parameters set to zero. The WER is only slightly different if no word class discounts are used. This can be explained by the fact that MT knowledge in form of mono-gram discounts is already optimally used by the cache LM. Though $w_{TM} = 0.075$ is comparatively low the discriminative capabilities of the TM lead to a further reduction in WER.

Technique	WER
Baseline ASR	12.6
Vocabulary Restrictions	$>$ 15.0
LM Interpolation	11.6
Hypothesis Selection (on Baseline)	10.5
Cache LM	10.4
Cache & Interpolated LM	10.1
Hypothesis Selection on Cache & Interp. LM	9.7
Hypothesis Selection on Cache LM	9.4

Table 1: Comparison of ASR improvement techniques

Hypothesis Selection on Cache & Interpolated LM System Output: When performing the hypothesis selection on the cache and interpolated LM system output we achieved a WER of 9.7% for w_{TM}=0.12, w_{LM}=0.15, sd=2.5, fp'=-10, lp'=5, n_{ASR}=150, n_{MT}=1 and all other parameters zero. The difference in WER towards rescoring on cache LM system output is insignificant.

2.4. Experiments and Results on MT component

For these experiments the n-best lists produced by the ”Hypothesis Selection on Cache LM” system were used. The experimental results are summarized in Table 2.

2.4.1. Language Model Interpolation

When interpolating the baseline LM with a small LM computed over the ASR n-best list, the best BLEU score, 53.4, was found for n=3 and an interpolation weight of i=0.8 for the small LM.

2.4.2. Retraining of the MT system

The ASR n-best lists were added several (x) times to the original training data and new IBM1 lexica (forward and backward lexicon) were computed. Two sets of experiments were run: the first with the translation memory fixed to the original training data and the second with the translation memory computed over the complete training data. In both cases a maximal BLEU score of 42.1, 70.2 respectively, could be found for the parameters n=1 and x=4.

2.4.3. Combination of LM Interpolation and Retraining

The above described systems for LM interpolation and retraining were combined. The best parameter settings were n=1, i=0.9 for LM interpolation and n=1, x=1 for retraining, yielding a BLEU score of 54.2, and 84.7 respectively.

3. Document Driven MTE-ASR System

Based on the results presented in chapter 2 we examined different combinations of the ASR and MT improvement techniques for our iterative MTE-ASR system design.

3.1. Data Set for MTE-ASR System Evaluation

The data set used for evaluating the iterative MTE-ASR system consists of 500 English and Spanish sentences in form and

	NIST	BLEU
Baseline MT	7.13	40.4
LM Interp	8.25	53.4
Update Translation Memory		
- Retraining	9.93	70.2
- Combination	10.90	84.7
Fixed Translation Memory		
- Retraining	7.28	42.1
- Combination	8.40	54.2

Table 2: Comparison of MT improvement techniques

content close to the BTEC. The English sentences were read 4 times, each time by 5 different speakers with 10 speakers overall. The data was split into four parts so that each sentence occurred just once per subset. Overall there were four MTE-ASR systems, one per subset. One tenth of each subset was randomly selected as held out data for tuning the parameters of the respective MTE-ASR system. The final performance was measured over the complete output of all four systems. Because of some flawed recordings the reduced data set consisted only of 1,747 sentences composed of 13,398 (959 different) words. The audio data equals 68 min.

3.2. Experiments and Results for the MTE-ASR System

The same ASR and MT baseline systems were used as in section 2. The OOV rate of the ASR system on the new data set was 0.51%. The perplexity of the language model used by both system was now 85.2 and thereby approximately four times higher than on the data set used for component evaluation. For ASR improvement, the cache LM approach as well as the mentioned combined techniques were taken into consideration. For MT improvement, the combination of LM interpolation and retraining was chosen, on the one hand with a fixed translation memory and on the other hand with an updated memory. The motivation for this was that, although the MT system with the updated memory yielded a much higher performance, complementary MT knowledge is lost by using it. The updated memory sees to it that primarily the ASR hypotheses added to the training data are selected as translation hypotheses.

For improving the ASR component, the combination of rescoring and cache LM in iteration 0 and the combination of rescoring, cache LM and interpolated LM in higher iterations yielded the best results. The better performance resulting from the additional use of LM interpolation after iteration 0 is due to the improved MT context information. For MT improvement it turned out that it is better to work with a fixed translation memory. The final WER was 1% absolute worse with the updated translation memory. No siginificant change in recognition accuracy was observed for iterations > 1. This was true for all examined system combinations that applied a subsequent rescoring on the ASR system output. If no rescoring was used, similar results to the case where rescoring was used could be obtained, but only after several (> 3) iterations. Figure 2 gives an overview on the components of our final iterative system design along with the respective performance values.

4. Conclusions

In this paper we examined several approaches for improving the ASR performance on the target language speech for human-mediated translation scenarios by incorporating all available knowledge sources in both the target and source language. The iterative system design that we developed from our experiments gave an reduction in WER of 37.6% relative. Attention should be paid to the fact that, even though the relative reduction in WER for iteration 0 was already very high, another significant improvement could be accomplished in iteration 1. We will extend our research for the non-document driven case in the future and will analyze the potential for improving the MT component in more depth.

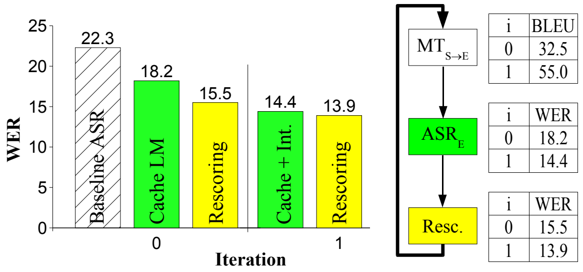

Figure 2: Final document driven system design; performance of the involved system components after iteration 0 and 1.

5. Acknowledgments

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). The authors would like to thank Matthias Eck for his help on the MT aspect of this work.

6. References

[1] M. Dymetman, J. Brousseaux, G. Foster, P. Isabelle, Y. Normandin, and P. Plamondon, “Towards an Automatic Dictation System for Translators: the TransTalk Project”, in *ICSLP*, Yokohama, Japan, 1994.

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

[3] P. Brown, S. Chen, S. Della Pietra, V. Della Pietra, S. Kehler, and R. Mercer, “Automatic Speech Recognition in Machine Aided Translation”, in *Computer Speech and Language, 8, 1994*

[4] P. Placeway, and J. Lafferty, ”Cheating with Imperfect Transcripts”, in *ICSLP*, Philadelphia, PA, USA, 1996

[5] Y. Ludovik, and R. Zacharski, ”MT and Topic-Based Techniques to Enhance Speech Recognition Systems for Professional Translators”, in *CoLing*, Saarbrücken, Germany, 2000

[6] H. Soltau, F. Metze, C. Fügen, and A. Waibel, ”A one-pass decoder based on polymorphic linguistic context assignment”, in *ASRU 2001*, Madonna di Campiglio, Italy, 2001

[7] F. Metze, Q. Jin, C. Fügen, K. Laskowski, Y. Pan, and T. Schultz ”Issues in Meeting Transcription - The ISL Meeting Transcription System”, in *ICSLP*, Jeju Island, Korea, 2004

[8] S. Vogel, S. Hewavitharana, M. Kolß, and A. Waibel, ”The ISL Statistical Machine Translation System for Spoken Language Translation”, in *IWSLT*, Kyoto, Japan, 2004