SPEECH RECOGNITION SYSTEM COMBINATION FOR MACHINE TRANSLATION

M.J.F. Gales[†], X. Liu[†], R. Sinha[†], P.C. Woodland[†], K. Yu[†], S. Matsoukas[*], T. Ng[*], K. Nguyen[*], L. Nguyen[*], J-L Gauvain[γ], L. Lamel[γ] and A. Messaoudi[γ]

[†]Cambridge University, Trumpington St., Cambridge, UK,
[*] BBN Technologies, Cambridge, MA, USA,
[γ] LIMSI-CNRS, Orsay Cedex, France

ABSTRACT

The majority of state-of-the-art speech recognition systems make use of system combination. The combination approaches adopted have traditionally been tuned to minimising Word Error Rates (WERs). In recent years there has been growing interest in taking the output from speech recognition systems in one language and translating it into another. This paper investigates the use of cross-site combination approaches in terms of both WER and impact on translation performance. In addition the stages involved in modifying the output from a Speech-to-Text (STT) system to be suitable for translation are described. Two source languages, Mandarin and Arabic, are recognised and then translated using a phrase-based statistical machine translation system into English. Performance of individual systems and cross-site combination using cross-adaptation and ROVER are given. Results show that the best STT combination scheme in terms of WER is not necessarily the most appropriate when translating speech.

Index Terms— Machine Translation, Speech Recognition

1. INTRODUCTION

The use of system combination approaches for speech recognition is now common in state-of-the-art speech recognition systems. There are two types of approaches to combining Speech-to-Text (STT) systems. The output from one system may be used as the adaptation supervision for a second system. This is referred to as *cross-adaptation*. The information propagated between the two systems is the 1-best hypothesis with associated confidence scores. The other approach is to take multiple system hypotheses and combine them together, for example using ROVER [1] or Confusion Network Combination (CNC) [2]. Both of these types of system combination have shown significant gains for reducing word error rates (WERs) both individually and conjointly. The gains from these approaches are increased when the cross-site combination is used, where the systems are built at different sites. This tends to result in greater diversity between the hypotheses than when using systems from a single site. In this paper, the impact of the form of cross-site STT combination scheme on speech translation is examined.

When combining multiple STT systems together for speech recognition the criterion commonly minimised is WER, or in the case of Mandarin Character Error Rate (CER). This is not necessarily the appropriate criterion to minimise if the output of the system is to be fed into a statistical machine translation (SMT) system. In particular as the majority of translation systems use phrase-based translation, it is important to maintain phrases and appropriate word-order. The choice of STT system combination approach for speech translation must take this into consideration. This issue is investigated for Arabic/English and Mandarin/English speech-to-text translation.

This work was supported in part under the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-C-0022.

2. STT SYSTEM COMBINATION

This section briefly discusses the two forms of STT system combination considered in this work. In particular issues that may impact the performance of SMT systems are discussed.

2.1. Hypothesis Combination

The first stage in most hypothesis combination schemes is to align the hypotheses against some "base" hypothesis. For ROVER combination [1] this is performed by minimising the Levenshtein distance, in CNC [2] the expected Levenshtein distance over the confusion networks is minimised. Once the hypotheses have all been aligned they are converted into a consensus network. The final output for each set of arcs is then made based on a combination of voting and confidence scores from the hypotheses. There are a number of issues to consider if the STT output is to be used for translation:

Alignment: the first stage in combining hypotheses together is to align the 1-best hypotheses or confusion networks against one another. As this alignment process is related to minimising a WER-like measure it can result in a number of peculiarities that may disrupt translation. For example the penalty from incorrectly inserting a word into a phrase is `1` for the recognition system. However, since it may break a possible phrase that could be translated in its entirety, the cost in terms of translation score may be far higher. This alignment issue becomes far more important the greater the difference between the hypotheses being aligned.

Voting/Selection: the criterion used to select the output from the consensus networks is tuned to minimise WER/CER. In contrast for translation Translation Edit Rates (TER) [3] or BLEU [4] scores are commonly used. As words/characters in the consensus network are selected independently of each other (given the aligned hypotheses), there are no phrase constraints imposed. Though this does not impact WER/CER, it may impact the translation performance. For schemes such as ROVER or CNC the selection may also make use of word-level confidence scores. In the case of combining two systems, confidence scores must be used. As the performance difference between STT systems increases, it becomes increasingly important to have good confidence scores. The use of poor word confidence estimates may result in inappropriate selections from the consensus network, again possibly breaking phrases.

2.2. Cross Adaptation

The vast majority of STT systems make use of an adaptation stage, normally based on linear transformations such as MLLR or CMLLR. To estimate these transforms an initial hypothesis is required. Normally this is obtained from an initial unadapted recognition run. However in *cross-adaptation* it is obtained from the output of another STT system, which may itself have made use of adaptation. Provided the adaptation transforms are not overly tuned to the adaptation supervision, gains are possible provided the errors in the supervision differ from those of the system to be adapted.

In terms of STT combination for translation, cross-adaptation may be viewed as a "safe" option. The system output does not involve the combination of two highly diverse systems. Note that there is still a level of hypothesis combination internal to each individual STT system described in this paper. However, in contrast to the cross-site hypothesis combination, the within-site hypotheses tend to be more consistent so shouldn't be severely impacted by the issues described in section 2.1[1].

3. STT POST PROCESSING

In processing data such as Broadcast News (BN) or Broadcast Conversations (BCs) for an STT system, the first stage is to segment the data into homogeneous blocks, or turns. Each block should consist of data belonging to an individual speaker in a particular acoustic environment. However, SMT systems are trained on text data, where individual sentences are aligned with each other and the punctuation is provided. It is therefore necessary to post-process the STT output so that "sentence-like" segments with punctuation are passed to the translation system. In this work different post-processing pipelines were applied to the Arabic and Mandarin data.

The Arabic STT output, whether generated by a stand-alone system or by combining multiple system outputs, was first divided into segments based on BBN's acoustic segmentation. This segmentation starts by detecting speaker turns in the audio, and then divides long speaker turns into shorter segments by splitting on pauses of significant duration (0.3 sec minimum). The resulting segments are about 10 sec long on average. The acoustic segmentation is further refined based on a Hidden-Event Language Model (HELM) [5], a Kneser-Ney (KN) smoothed 4-gram trained on 850M words of Arabic news. The HELM integrates pause duration as observation into the HMM search, and makes use of a bias to insert boundaries at a high rate, then boundaries of low posterior probability are removed while constraining the maximum sentence length to 51 words.

For the Mandarin system the post-processing consisted of a two-stage process. First a simple silent splitting process based on the inter-word silence gaps was performed. The data was split at silences of greater than 0.9 seconds. In addition a maximum segment length of 20 seconds was imposed by splitting longer segments at the longest inter-word gap. A HELM was then used to further split the segments.

In addition to adding sentence boundaries and sentence end punctuation, both STT outputs had a *reverse number mapping* applied. STT systems are built on acoustic data where numbers are in their spoken form. However, translation systems are built on text data, where numbers are normally written in digit form. Language specific number mappings were generated, and applied prior to translation.

[1] In preliminary evaluation on the Mandarin, there was a slight advantage in combining, using CNC, multiple branches from within the CU system.

4. RESULTS

The experiments described in this section made use of STT systems built at three different sites, BBN, Cambridge University (CU) and LIMSI. The final, sentence boundary marked, STT output was translated using an SMT system developed at BBN.

The BBN translation engine employs statistical phrase-based translation models, with a decoding strategy similar to [6]. Phrase translations are extracted from word alignments obtained by running GIZA++ [7] on a bilingual parallel training corpus (139M words of Arabic/English and 223M words of Chinese/English). A significant portion of the phrase translations are generalized through the use of part of speech classes, for improved performance on unseen data. Both forward and backward phrase translation probabilities are estimated and used in decoding along with a pruned trigram English LM, a penalty for phrase reordering, a phrase segmentation score, and a word insertion penalty. These scores are combined in log-linear fashion, with weights estimated discriminatively, on held-out BN and BC data in order to minimize overall TER [3]. The output of the decoding process is an N-best list of unique translation hypotheses, which is subsequently rescored using an unpruned, KN smoothed 5-gram English LM. This LM is trained on more than 5B words of text, consisting of the Gigaword corpus, archives downloaded from news web sites, and other news and conversational web data. The majority of the data used for training the SMT system was text, not speech. The translation systems is thus not ideally matched to translating speech. However, it should be more closely matched in style to BN data, rather than the more conversational BC data.

As the translation systems were tuned for TER, translation performance is assessed in this work in terms of TER. Note, the same general trends were observed with BLEU, though the differences between the systems were reduced. For these experiments only a single translation reference was available.

4.1. Arabic

Two STT systems, from BBN and LIMSI, were evaluated both stand-alone and in combination. The individual system descriptions are below.

LIMSI: The LIMSI Arabic STT system used in the GALE'06 evaluation uses the same basic modeling and decoding strategy as described in [8]. Word recognition is performed in three passes, where each decoding pass generates a word lattice with cross-word, position-dependent, gender-dependent Acoustic Models (AMs), followed by consensus [9] decoding with 4-gram and pronunciation probabilities. Unsupervised AM adaptation is performed for each segment cluster using the CMLLR and MLLR techniques and relies on a tree organization of the tied states to create the regression classes as a function of the available data. Different combinations of automatic segmentations (GMM or BIC based), acoustic models and language models are used in the different passes [10]. The acoustic models are MLLT-SAT trained on 1044 hours of data. Short vowels are modeled explicitly in the system and a generic vowel was introduced to enable training on non-vocalized data. A small model set with 5.1k tied states (64 Gaussians per state) is used in the first pass. Later passes use models covering 27.4k phone contexts with 11.5k tied states. N-gram language models were trained on a total 852M words of texts including transcripts of the audio data. The 4-gram model is interpolated with a neural network language model trained on the acoustic transcriptions and a subset of the text data, and used to rescore the lattices after the last two passes.

BBN: The BBN Arabic STT system uses a similar modeling and

search strategy as described in [11]. The multi-pass recognizer first does a fast match of the data to produce scores for numerous word endings (word graphs) using a coarse state-tied mixture acoustic model (AM) and a bigram language model (LM). Next, a state-clustered tied-mixture (SCTM) AM and a trigram LM are used to decode the word graphs to produce lattices. The lattices are then rescored using a cross-word SCTM AM and a 4- gram LM. The best path of the rescored lattice is the recognition results. The decoding process is repeated two (or three) times with speaker-independent AMs used in the first stage while subsequent decoding stages use speaker-adaptively-trained AMs. All AMs were trained on about 1300 hours of speech data with the largest model having about 6k states and 800k Gaussians. All LMs were estimated based on a training corpus of almost 1 billion words.

Cross-adaptation in Arabic was run in both directions as neither of the systems was clearly better than the other on all the data. For both directions of cross-adaptation, no use was made of adaptation supervision confidence scores.

Two test sets were used: bnad06, consisting of about 3 hours of BN data, collected in November 2005 and January 2006; and bcad06 comprised of about 3 hours of BC data, collected in the same epoch.

System	WER%	
	bnad06	bcad06
BBN	18.9	29.4
LIMSI	19.7	28.8
LIMSI→BBN	18.3	28.9
BBN→LIMSI	18.4	27.6
BBN⊕LIMSI	17.7	27.4

Table 1. STT performance (WER %) comparing baseline, cross-adaptation (LIMSI→BBN) and ROVER combination (BBN⊕LIMSI)

Table 1 shows the performance of the individual systems, cross-adaptation in both directions, and ROVER combination. For the BN test set, cross-adaptation in either direction gave gains of about 0.5% absolute over the best individual system. The use of ROVER combination gave an additional 0.6-0.7% absolute gain over cross-adaptation. On the BC data the use of cross-adaptation LIMSI→BBN yielded a slight degradation in performance. Again the best performing system in terms of WER was ROVER on BC data.

System	bnad06		bcad06	
	lc	mix	lc	mix
BBN	62.20	64.57	67.86	69.80
LIMSI	62.57	64.83	67.23	69.22
LIMSI→BBN	62.13	64.57	67.78	69.83
BBN→LIMSI	61.89	64.27	67.14	69.20
BBN⊕LIMSI	62.18	64.43	67.02	69.16

Table 2. Lower-cased (lc) and mixed (mix) TER (%) performance of baseline, cross-adaptation (LIMSI→BBN) and ROVER (BBN⊕LIMSI)

Table 2 shows the TER scores, both lower-cased and mixed, for the STT system outputs from Table 1. For the BN data, though the lowest WER was obtained with ROVER combination, the lowest TER score was with the BBN→LIMSI cross-adaptation system. Though the BBN and BBN⊕LIMSI systems differed by 1.2% absolute in WER there was minimal difference in TER. Worse translation performance is observed on the BC data compared to the BN data. This is expected as this data is both harder, being more conversational in style, and there is less well represented in the training data for the translation system. On this data, there is little difference in translation performance between the LIMSI, BBN→LIMSI and BBN⊕LIMSI STT configurations, despite a difference of over 1.0% absolute in terms of WER.

4.2. Mandarin

Two STT systems were constructed and combined for the Mandarin BN and BC transcription tasks. Both systems were trained on around 503 hours of acoustic data, 156 hours of BC data and 347 hours of BN data. The language models for each of the systems were trained on over a billion words of text data, which included the acoustic data transcriptions, broadcast transcriptions and news-wire data. The details of the two system are summarised below.

BBN:The underlying technologies of the BBN Mandarin STT system are essentially the same as those used in the Arabic STT systems described above. In addition, pitch features are used to accommodate the *tonal* phonemes existing in Mandarin (as described in detail in [12]).

CU: The overall structure of the CU system was similar to that described in [13]. This comprises an initial lattice generation stage followed by lattice rescoring with a 4-gram language model and adapted acoustic models. In this work two forms of acoustic models were combined. Both systems were based on PLP with 1st, 2nd and third derivatives using HLDA to project to 39 dimensions with CMN and Gaussianisation. The first was a gender-dependent system, the second used speaker adaptive training. Initially standard 1-best adaptation was performed followed by lattice-based adaptation with multiple speech baseclasses. The two models were combined using CNC.

For cross adaptation the BBN output was used as the supervision to adapt the models in the lattice rescoring stage of the CU system. Rather than using lattice-based adaptation, confidence-based adaptation using the confidence level scores from the BBN system was run. In Mandarin there is an additional problem that must be considered. Most Mandarin STT systems use language models based on words. However the majority of text is not split into words, but consists of sequences of characters. Thus the first stage in training a Mandarin STT system is to run a character-to-word (C2W) segmenter on the training text. Though both the BBN and CU STT systems used the same algorithm, a longest first match, different multi-character word-lists were used. Thus the C2W segmentations differed. In terms of hypotheses combination there are two possible levels to operate at. First the systems may be combined at the character level, which may make any alignment problems more severe, but is consistent with the STT CER criterion. Alternatively they may be combined at the word level. This requires resegmenting the outputs to be consistent, which may yield strange word sequences for errorful hypotheses. Both forms are investigated in this section.

Two test sets were used to evaluate the performance of the systems. The first, bnmdev06, comprises 3.6 hours of BN data. This was taken from 12 shows and included the RT04f Mandarin evaluation test set and the mainland shows from the 2003 evaluation test set. The first half of this test set was used for tuning the translation

system, so results are only quoted on 1.8 hours set referred to as bnmd06. bcmdr06 consists of snippets from a range of BC shows yielding a total of 0.31 hours of data.

System	Comb. level	CER%	
		bnmd06	bcmdr06
CU	—	8.0	21.8
BBN		8.6	23.9
BBN→CU	—	7.5	20.7
CU⊕BBN	`char`	7.1	21.3
	`word`	7.8	21.4

Table 3. STT performance (CER %) comparing baseline, cross-adaptation (BBN→CU) and ROVER combination (CU⊕BBN)

Table 3 shows the performance in terms of character error rate (CER) of the various systems. The CU baseline system is approximately 6-7% relative better than the BBN system on both BN and BC data. However by using the BBN system output for supervision in cross-adaptation the CER was further reduced by 0.5% absolute, 7% relative, on the BN data and 1.1% absolute, 5% relative, on the BC data. The best CER performance was obtained using character level ROVER combination. The C2W resegmentation associated with the word-level ROVER appears to have had an impact in terms of CER, degrading performance by 0.7% absolute on BN and 0.1% on BC compared to character-level ROVER.

System	Comb. level	bnmd06		bcmdr06	
		lc	mix	lc	mix
CU	—	67.94	70.00	75.00	76.94
BBN→CU	—	67.28	69.33	74.51	76.58
CU⊕BBN	`char`	67.41	69.50	75.19	77.07
	`word`	67.60	69.74	75.47	77.37

Table 4. Lower-cased (lc) and mixed (mix) TER (%) performance of baseline, cross-adaptation (BBN→CU) and ROVER (CU⊕BBN)

Table 4 shows the lower-cased and mixed TER scores for the CU and combined systems from table 3. For the BN data, in common with the Arabic system, cross-adaptation outperformed ROVER combination (at the character level) despite having a higher CER. As previously discussed the alignment issues may be more severe for character level ROVER than word-level combination. For the BC data, where the lowest CER system used cross-adaptation, the lowest TER was also the cross-adapted system.

5. CONCLUSIONS

STT system combination is a standard approach in state-of-the-art speech recognition systems. Typically these combination approaches are tuned to minimising WER, or CER. If the output of the STT process is to be translated then the appropriate metric to consider is the translation performance. This paper has examined the impact of the form of STT system combination has on the translation performance. Two forms of STT system combination were examined, cross-adaptation and ROVER hypotheses combination. The performance of speech translation was examined in two different language pairs, first Arabic/English, second Mandarin/English. For each language pair both BN and BC data was recognised and translated.

For the BN data ROVER system combination was found to yield the lowest error rate for the STT systems in both Mandarin and Arabic. However this "better" STT output was not reflected in improved translation performance. This may indicate that some of the alignment issues discussed in section 2 are impacting the phrase-based translation system. Unfortunately given the current performance of the translation system it is hard to clearly identify regions where this is occurring. For the BC data the trends are less. This may be due to overall higher error rates and the lack of appropriate training data for the MT system. Overall, though ROVER combination usually gave the lowest error, the use of cross-adaptation was generally found to be a safer STT combination scheme for translation.

6. REFERENCES

[1] J.G. Fiscus, "A post-processing system to yield reduced word error rates: Recogniser output voting error reduction (ROVER)," *Proc. IEEE ASRU Workshop*, pp. 347–352, 1997.

[2] G. Evermann and P. C. Woodland, "Posterior probability decoding, confidence estimation and system combination," in *Proceedings Speech Transcription Workshop*, College Park, MD, 2000.

[3] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in *Proceedings of Association for Machine Translation in the Americas*, 2006.

[4] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: a method for automatic evaluation of machine translation," Tech. Rep. RC22176 (W0109-022), IBM Research Division, 2001.

[5] A. Stolcke and E. Shriberg, "Automatic linguistic segmentation of conversational speech," in *ICASSP*, 1996.

[6] P. Koehn, F. Och, and D. Marcu, "Statistical phrase-based translation," in *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference*, 2003.

[7] F. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[8] J.L. Gauvain, L. Lamel, and G. Adda, "The LIMSI broadcast news transcription system," *Speech Communication*, May 2002.

[9] L. Mangu, E. Brill, and A. Stolke, "Finding consensus among words: Lattice-based word error minimization," in *Proceedings Eurospeech'99*, 1999.

[10] S. Matsoukas et al, "Advances in transcription of broadcast news and conversational telephone speech within the combined EARS BBN/LIMSI system," *IEEE Trans. on Audio, Speech and Language Processing*, September 2006.

[11] L. Nguyen, B. Xiang, M. Afify, S. Abdou, S. Matsoukas, R. Schwartz, and J. Makhoul, "The BBN RT04 English broadcast news transcription system," in *Proceedings InterSpeech*, Lisbon, Portugal, September 2005.

[12] B. Xiang, L. Nguyen, X. Guo, and D. Xu, "The BBN Mandarin broadcast news transcription system," in *Proceedings InterSpeech*, Lisbon, Portugal, September 2005.

[13] R. Sinha, M. J. F. Gales, D. Y. Kim, X. A. Liu, K. C.Sim, and P. C. Woodland, "The CU-HTK Mandarin broadcast news transcription system," in *ICASSP*, 2006.