ARABIC ASR AND MT INTEGRATION FOR GALE

Yaser Al-Onaizan, Lidia Mangu

IBM T.J. Watson Research Center
1101 Kitchawan Road, Route 134, Yorktown Heights, NY 10523

ABSTRACT

In this paper we describe our work in machine translation of Arabic speech into English. This work was done within the context of the GALE research program. We describe several integration techniques between our ASR and MT system. Our initial results suggest that tighter coupling between ASR and MT system improves the translation quality of speech input. We explore the effect of each integration technique on the overall system.

Index Terms— Speech Translation, Statistical Machine Translation, Arabic Spoken Language Translation, GALE Translation System

1. INTRODUCTION

Automatic Speech Recognition (ASR) and Statistical Machine Translation (SMT) systems are often developed independently of each other. In order to build a speech translation systems, researcher typically glue ASR and SMT systems as two black boxes where the output of the ASR system is fed to the SMT system. A tighter integration is often necessary to improve the overall accuracy of the translation.

Most of the work on tight integration of ASR and SMT systems, thus far, have focused on limited vocabulary and limited domain translation (e.g., travel and emergency medical diagnosis). Such research are mostly conducted in two major research programs The Verbmobil research program in Europe [1] and CAST (formerly Babylon) research program sponsored by the Defense Advanced Research Projects Agency (DARPA). Such systems are typically have limited recognition and translation vocabulary. The utterances are short and have very simple sentence structure (e.g., ...). However, The demand for large vocabulary news domain speech translation have dramatically increased with the explosion in the number of accessible TV and Radio stations in foreign languages. Early attempts to address this new challenge are driven by two recent research programs: TC-STAR and GALE.

GALE is a research program sponsored by DARPA. Speech translation is a major part of the GALE program, where Arabic and Chinese broadcast news and broadcast conversations are translated into English. This presents many serious challenges to both ASR and SMT systems.

2. DESCRIPTION OF OUR SPEECH TRANSLATION SYSTEM

2.1. Speech Recognition System

The speech recognition system used in these experiments has two components, one that explicitly models the short vowels, which are pronounced in Arabic but almost never transcribed (vowelized system), and one that does not (unvowelized system). The final system is a cross-adapted system as the transcripts generated by the unvowelized system were used to train the speaker-adapted transforms for the vowelized one. Both intermediate systems use a pentaphone acoustic context, 5K context dependent states and 400K 40-dimensional diagonal-covariance Gaussians. They are trained using a fMPE and MPE on 135 hours of unsupervised data and 1800 hours of TDT-4 BN03 unsupervised data. The language model is a 617K vocabulary 4-gram LM with 56M n-grams. The ASR system is described in more details in [2].

2.2. Statistical Machine Translation System

The SMT system used in these experiments is a phrase-based SMT system. The SMT system is developed in two stages. The training phase in which two components are trained offline. The second phase is the decoding phase, where a new novel sentence is translated by the system by searching among thousands of possible translations and choosing the won that minimizes a log-linear cost function. The hypothesis translations are generated based on the statistical models built during the training phase. The log-linear function is defined in terms of several statistical models such as the phrase-table, the n-gram language model, and so on. More details about this can be found in [3].

The SMT system is trained using two types of data. The first type is bilingual and is used to train the phrase-pair set and the probabilistic translation lexicons and other statistical models. These are typically trained on parallel data where source sentence (e.g., Arabic) are paired with their translation (e.g., English). The other type monolingual, in which text in the target language (e.g., English) is used to train an n-gram language model. SMT systems require large amounts of text to train them, typically in the order of 100 million words of parallel text and billions of words of monolingual

text. However, there is very limited speech data to train SMT systems. Most of the data available from SMT training is of the text genre (e.g., newswire, UN proceedings, Parliamentary proceedings, etc.) Therefore, one must pay special attention when integrating ASR and MT systems into a single overall speech translation system in order to minimize the mismatch between them.

The parallel data used to train the IBM 2006 GALE SMT system used in these experiments is composed of 6.5 million source words of news, 200k source words of speech transcripts, and 115 million source words on UN proceedings. The English language model is trained on 3 billion words of English news from the LDC's English Gigaword Second Edition Corpus (Catalog Number LDC2005T12).

In the following section, we describe several techniques that tightly integrates our ASR and SMT system in order to alleviate potential mismatches between what the ASR system produces as output and what the SMT system expects as input.

3. ASR AND SMT SYSTEM INTEGRATION

We describe several techniques that tightly integrate our ASR and SMT system in order to alleviate potential mismatches between what the ASR system produces as output and what the SMT system expects as input. Also, historically ASR systems are optimized using word error rate as the objective function; when the output of the speech recognizer is being fed through a machine translation system, then the WER becomes only one of the dimensions to be optimized. We present a number of experiments for testing the sensitivity of machine translation system to the speech recognizer output.

3.1. Vocabulary and Phrase Integration

The initial ASR vocabulary contained only words from all the data used for language modeling. The first step towards better ASR-MT integration was the addition of all the words from the MT vocabulary which occur at least twice in the MT training data. The OOV rate of the final 617K vocabulary on a variety of test sets is below 0.8%. The next step is to pay attention to the phrase table used in the MT system, and add these phrases as compound words in the ASR vocabulary. The number of entries in the MT phrase table is prohibitively large, therefore we extracted only the ones which occur 20 times or more, whose word components exist already in the ASR vocabulary. A new language model is trained on a corpus created by replacing sequences of words with the corresponding phrases from the newly created 800K vocabulary. The usefulness of the new phrases in the vocabulary is assessed in a lattice rescoring framework. The word lattices built using the 617K vocabulary are converted into phrase lattices by replacing sequences of word arcs with one arc bearing the corresponding phrase. The new lattices are rescored used the phrase LM described above and the results are shown in Table 1. Notice that there is a small WER improvement even though the motivation for having a phrase vocabulary was to improve the translation output.

Hyp	BNAT05	BCAD05
617K w LM	15.3%	23.1%
800K w+ph LM	15.2%	22.9%

Table 1. Word Error Rates before and after rescoring with a phrase-based LM

System	BNAT05	BCAD05
617K w LM		
800K w+ph LM		

Table 2. TER and BLEU results for the phrase based system compared to the original word based

The effect on the translation quality can be seen in Table 2.

In the future we intend to use the 800K vocabulary as the main recognition vocabulary. The challenge of this approach is generating baseforms for long phrases, especially for the vowelized system in which the number of pronunciations per word is much higher.

3.2. Deletions/Insertions Ratio

The output of a speech recognition system is the hypothesis with the highest score, commonly computed by combining the acoustic model score and a weighted language model score. By varying the weight of the language model we can obtain outputs with completely different ratios of deletions and insertions, even when the overall WER is almost the same. Then the question is which alternative is better for translation, the assumption being that it should affect the translation quality. But Table 3 shows that the MT performance in terms of TER and BLEU is almost unchanged even when the ratio of deletions and insertions is 3 times higher.

3.3. Sentence Segmentation

The first step of a speech recognition process is the segmentation of the audio into speech and non-speech segments. The input to the MT system can be either segmented according

WER	ratio del/ins	BLEU	TER
18.3	9	17.80	66.76
18.2	3.5	17.86	67.35

Table 3. Sensitivity of TER and BLEU on the ratio Del/Ins of the speech output

Segmentation	BLEU	TER
auto	16.40	65.0
auto + periods	17.51	64.24
ref	16.67	64.66
ref + periods	17.85	63.84

Table 4. Sensitivity of TER and BLEU on the segmentation of the speech output

to this initial segmentation or resegmented using a sentence end detection algorithm. In order to assess the importance of using the later, we concatenated the output of the recognizer and resegmented it according to the reference segmentation. Table 4 shows the TER and BLEU scores for the two different segmentations. It also shows the scores when we artificially attach a period mark at the end of each segment from either automatic or reference segmentation. The conclusion is that the speech/non-speech segmentation is very close to the reference segmentation in terms of translation quality. An even more important conclusion is that we should add a period at the end of each segment, no matter which segmentation method we choose.

3.4. Punctuation Insertion

ASR systems typically produce text output with no punctuations (e.g., comma, period, etc.). However, as we described in Section 2.2, SMT systems are typically trained on text genre where punctuations are very common. This mismatch potentially reduces the number of phrase-pairs in the SMT's phrase table that match a given source sentence. This in turn severely restricts the number of matches explored by the SMT decoder and hence may miss on some potentially good translations. Additionally, it is desirable for SMT output to be fluent and hence having the right punctuation in the output is required. In some cases, having the right punctuation is actually crucial to reflect the intended meaning of the source sentences.

An ASR system's output lack of punctuation can be addressed in three different approaches. The first approach is to insert punctuations in the ASR output after the recognition phase but before feeding it into the MT system. Several research groups have worked in insert punctuation in English ASR output such as. However, we are not aware of any work in inserting punctuations in Arabic ASR output. The second approach is to insert punctuations in the MT output as described in [4]. This approach addresses the fluency requirement of the output. However, it does not alleviate the potential mismatch between the ASR output and the MT phrase table. The third approach is to modify the SMT system such that it tolerates the lack of punctuation in its input but able to produce punctuation in its output. The latter approach is what we adopt in this paper. Our SMT system is modified such that punctuations are ignored when looking up a source phrase from the source sentence in our phrase table. It is important to note that punctuations in the target part of the phrase table are left intact and hence will be produced by the SMT system. This punctuation-ignoring modification can be achieved by modifying the phrase table itself without the need to modify the SMT system itself. This is done by removing punctuations from the source side of every phrase pair in our phrase table. For example, ...

It is important to delay this punctuation removal after extracting the phrase table from the training data. Removing the source punctuations before word-aligning the training data might affect the accuracy of the alignment algorithm since target punctuations will not have anything to align to in the source.

3.5. Confusion Network Translation

In addition to producing the usual 1-best output, our ASR system is capable of producing a confusion network. A confusion network is a sausage-like lattice with many alternatives produced at each node. We modified our SMT system to accept confusion networks when they are present. In essence, a 1-best ASR output where at each node there is only one alternative at each position. In case of a 1-best ASR output, the SMT system extracts all phrase pairs that match the source ngrams (adjacent phrases) in the input sentence. When a confusion network is given as input to the SMT system, the SMT system extracts all phrase pairs that match any paths in the confusion network.

Our preliminary results show insignificant gains from translating confusion networks over 1-best output. We are in the process of investigating the lack gains from translating confusion networks. One method we are exploring to utilize alternatives in the confusion network is to weigh alternative by the weight of the path in the confusion network.

4. EXPERIMENTS

In this section, we describe several experiments that we conducted to investigate how some of the techniques we described in Section 3 affect the quality of the overall speech translation system as measured by BLEU [5] and TER [6].

4.1. Punctuation Insertion Experiment

The test set we use for this experiment is the NIST MT Eval 03 test set, which has 663 Arabic sentences. To simulate the speech output, we removed all punctuations from the test set source, but the reference translations are left unchanged. To study the effect of our punctuation insertion technique, we ran our SMT system with two sets of phrase tables. The first phrase table is our regular table extracted from parallel data. The second phrase table is a modified version of the first phrase table in which punctuations are removed from all

ASR WER	BLEU	TER
19.3	19.99	65.33
15.9	20.43	63.11
0	25.08	57.89

Table 5. Sensitivity of the SMT system TER and BLEU to the ASR word error rate

source phrases, but are left intact in the target side. When we decoded the MT03 test set using the first phrase table, we obtained a BLEU score of 41.78. When the altered phrase table is used, the BLEU score obtained is 43.62. The filtering technique we described shows a statistically significant increase in BLEU score.

4.2. Effect of ASR WER on Translation Quality

In this section, we discuss how varying the ASR word error rate affects the BLEU score of the translation. Table 5 shows BLEU and TER scores where the ASR WER is 19.3, 15.9, and 0. The 0 error rate is when the human transcripts are used as input to the MT system instead of using the ASR output. Significant reduction in WER rates lead to significant reduction in TER and significant increase in BLEU, but not at the same rate.

5. CONCLUSION

Large vocabulary, spoken language translation presents unique challenges to machine translation systems. We described in this paper several techniques to alleviate some of the challenges in integrating large vocabulary ASR and SMT systems. Some of the techniques described had significant improvements over translation quality.

6. ACKNOWLEDGMENT

This work was partially supported by the Defense Advanced Research Projects Agency GALE program under contract No. HR0011-06-2-0001.

7. REFERENCES

[1] Wolfgang Wahlster, *Verbmobil: Foundations of Speech-to-Speech Translation*, Springer, 2000.

[2] H. Soltau, G. Saon, D. Povey, L. Mangu, B. Kingsbury, J. Kuo., M. Omar, and G. Zweig, "The IBM 2006 GALE Arabic ASR System," in *Submitted to ICASP-2007*, 2007.

[3] Yaser Al-Onaizan and Kishore Papineni, "Distortion models for statistical machine translation," in *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia, July 2006, pp. 529–536, Association for Computational Linguistics.

[4] Young-Suk Lee, Yaser Al-Onaizan, Kishore Papineni, and Salim Roukos, "Ibm spoken language translation system," in *TC-STAR Workshop on Speech-to-Speech Translation*, Barcelona, Spain, June 2006, pp. 13–18.

[5] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "BLEU: a Method for Automatic Evaluation of machine translation," in *40th Annual Meeting of the Association for Computational Linguistics (ACL 02)*, Philadelphia, PA, July 2002, pp. 311–318.

[6] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul, "A study of translation edit rate with targeted human annotation," in *Proceedings of Association for Machine Translation in the Americas*, 2006.