Optimizing Sentence Segmentation for Spoken Language Translation

Sharath Rao, Ian Lane, Tanja Schultz

InterACT, Language Technologies Institute,
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

ABSTRACT

The conventional approach in text-based machine translation (MT) is to translate complete sentences, which are conveniently indicated by sentence boundary markers. However, since such boundary markers are not available for speech, new methods are required that define an optimal unit for translation. Our experimental results show that with a segment length optimized for a particular MT system, intra-sentence segmentation can improve translation performance (measured in BLEU) by up to 11% for Arabic Broadcast Conversation (BC) and 6% for Arabic Broadcast News (BN). We show that acoustic segmentation that minimizes Word Error Rate (WER) may not give the best translation performance. We improve upon it by automatically resegmenting the ASR output in a way that is optimized for translation and argue that it might be necessary for different stages of a Spoken Language Translation (SLT) system to define their own optimal units.

Index Terms— Automatic Speech Recognition, Statistical Machine Translation, Sentence Segmentation, Optimal segment length

1. INTRODUCTION

With significant growth in the performance of Automatic Speech Recognition (ASR) over the past two decades, new problems in language technologies are being pursued that use the output of an ASR system as the input for other applications. These applications include among others Spoken Language Translation systems (SLT), speech summarization and dialog systems. However, due to the spontaneous nature of spoken language, sentences are not well defined as in written text. Since most of these systems require structure in the ASR output stream, segmenting ASR output into sentence-like units is an intermediate step that can have significant impact on the overall performance of these systems.

Previous work in sentence segmentation has focused on spotting sentence boundaries as defined by humans where performance was typically evaluated in terms of precision/recall or Sentence Unit error rates [1], [2]. While such measures may be appropriate for rich transcription tasks, a system optimized to detect manually annotated sentence boundaries has not been shown to be optimal for speech translation. In planned speech such as lectures and broadcast news, sentences tend to be long and are often composed of syntactically and semantically independent units. Translating these long sentences as a single whole, in addition to being a huge computational burden might not be optimal.

Different motivations have guided previous work in sentence segmentation as a pre-processing step in translation tasks. In [3], a technique was proposed to efficiently use training data by splitting long training examples and improving model estimation for Statistical Machine Translation (SMT). Sentence splitting has been used to improve Example-Based Machine Translation (EBMT) performance where longer sentences do not yield good translations. [4] proposed a technique to split sentences by matching sentences to those in corpus using editing distance criterion and showed improvement in EBMT performance. However, no results on effects of recognition errors were reported. In [5], long sentences were split to reduce parsing complexity. The approach described in [6] splits sentences before and during parsing to improve translation performance for a Interlingua-based Spanish-English MT system. The above approaches, however, have focused on limited domain tasks and are not easily extendable to open domain problems such as translation of broadcast news and conversations.

The goal of this paper is to show that segmentation of ASR output has to be optimized taking into account the downstream process that will be applied. We investigate optimizing segmentation to improve translation accuracy and show that segmenting within known sentence boundaries can improve end-to-end performance.

2. MOTIVATION

To motivate the discussions in this paper, we report results from a pilot experiment where we translated transcriptions of 5 broadcast conversation shows by considering 2 different methods of translating sentences. In the first case, no segmentation was performed, effectively translating each complete sentence. In the second case, a segment boundary was marked at commas and periods.

Table 1. Effect of sentence segmentation using commas and periods in Broadcast Conversation transcripts

Segmentation type	Avg. segment length	BLEU
Complete sentence	18.4	17.33
Segment at every comma	9.9	20.49

Table 1 shows the difference in translation performance obtained from segmenting sentences before translation. Translating an entire

This work is 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.

sentence was found to result in a significantly lower BLEU score than when each sentence is segmented at a comma prior to translation. This suggests that locating commas in addition to periods helps define independently translatable regions within a sentence and results in improved translation. However, speech translation systems work on the output of an ASR system where no comma or period information is available. Moreover, the notion of punctuation for spoken language is unclear as evidenced in significant interannotator disagreement in such tasks [7]. These aspects in addition to the degradation in translation due to ASR errors together present a challenge for speech translation systems.

3. SENTENCE SEGMENTER

To perform sentence segmentation on ASR output, we used the approach followed in the ISL TC-STAR Spring 2006 evaluation system. A detailed description of this approach can be found here [9]. Pause duration at each word was obtained by computing the difference between start time of a particular word and end time of the previous word from the ASR first-best output. Since acoustic/prosodic features such as pitch and energy did not yield significant improvement over LM probabilities and pause duration, they were not used in this work.

Our experiments indicated that using the pause duration at each boundary to make a first pass decision before applying the LM helped in improving precision. Only those word boundaries whose corresponding pause lengths fell within a set range were considered as candidates for segment boundaries. The range of allowable pause duration was tuned on the development set. For these experiments, all boundaries with pause durations higher than 0.03 seconds and lower than 0.76 seconds were considered for LM scoring. Those lower than 0.03 seconds were hardcoded to be normal word boundaries whereas those above 0.76 seconds were marked as segment boundaries. Once the candidate segment boundaries are identified using the above criterion, the question of whether to segment or not is decided by the LM probability scores. A threshold γ on the ratio of log-likelihood of segment boundary to that of word boundary is used to control the average number of words per segment.

$$\delta = \frac{\text{Log-likelihood of segment boundary}}{\text{Log-likelihood of word boundary}} \quad (1)$$

if $\delta <= \gamma$, then sentence boundary else word boundary

4. EXPERIMENTAL SETUP

4.1. System description

For our ASR experiments, the ISL Arabic ASR system was used [8]. The MFCC-based acoustic model of the ASR system was trained on 190 hours of Arabic speech data of which broadcast conversation comprised 60 hours, with the rest being the broadcast news component. The language model was trained on the Arabic gigaword corpus with an additional small component containing broadcast conversation transcripts from the web. The output of the first-pass speaker independent decoding was used in all our experiments. The 4-gram language model used in the sentence segmenter was trained on 32 million words from the Arabic gigaword corpus.

For translation experiments, we used the Arabic-to-English phrase-based SMT system developed at the ISL [10]. This system was trained on 3.4 million sentences from the Arabic-English bilingual data comprising the UN data and news corpora provided by the LDC. The language model for this system was trained on the English side of the above data containing nearly 100 million words. The optimal alignment model combination parameters were obtained by performing Minimum Error Rate Training (MER) [11] on the development sets. Separate optimizations were performed for BN and BC shows. We report translation performance in terms of BLEU computed on a single reference translation (constrained by availability) per sentence [12]. The low BLEU scores (relative to the state-of-the-art systems) reported in this paper can be explained from the fact that only a single reference translation was available per sentence.

4.2. Datasets

We investigate the effect of segmentation on two standard datasets - one each for BN and BC shows. For BC, we use the BCAD05 dataset with 3 shows of 30 mins each designated as evaluation set (BCAD05-E) and 2 as development set (BCAD05-D). For experiments on BN data, we use 2 shows from the RT04 dataset as our development set and 5 shows from the BNAT05 dataset comprising 660 sentences as the evaluation set. Table 2 lists the details regarding specific shows used from the various datasets.

Table 2. Data and Showname listing

Genre	Development Set	Evaluation Set
Broadcast Conversation	ALJZ-2005-02-18 ALJZ-2005-02-22	ALJZ-2005-02-16 ALJZ-2005-03-01 ALJZ-2005-03-11
Broadcast News	ALJ-031208-060215 DUB-031211-113227	20010127-1100-VOA 20010129-1530-NTV 031124-113208-DUB 031122-133544-ALJ 031121-150530-LBC

5. RESULTS AND DISCUSSION

5.1. Improving translation through intra-sentence segmentation

First, manually segmented (human-defined sentences) audio data was decoded and the first-best ASR hypotheses were obtained. Translation performance for these hypotheses forms the baseline for comparing segmentation-translation performance. Next, using the above segmenter, these hypotheses are further segmented by varying γ in (1) to obtain different degrees of segmentations for each sentence. These segments were then translated independently using the ISL Arabic-English SMT system. To evaluate translation performance, for each sentence in the test set, a single translation hypothesis was formed by combining in the same order all the segment translations corresponding to that sentence. We performed the above experiments on the BN and BC development sets. Fig. 1 shows the translation performance with different segmentation for the 2 BN shows. We quantify different segmentations in terms of the average length of an input segment. With respect to the baseline, we see a steady improvement in BLEU score as the average segment length decreases. The BLEU score peaks when average segment length is about 8-9 words long, after which it drops sharply and translation performance suffers. The reason for this is that while too long segments result in heterogeneous phrases that are better translated separately, too short segments cause loss of context and hence degrade translation quality.

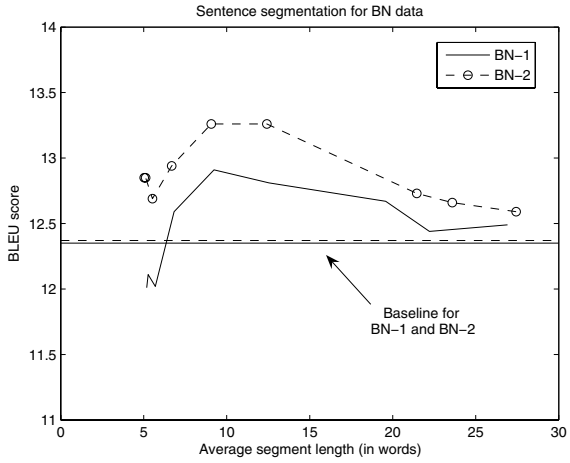

Fig. 1. Effect of segmentation on translation for 2 shows BN-1 and BN-2 from BN Dev. set; *Baseline performance (manual segmentation) for BN-1 and BN-2 is marked*

For both the shows, the optimal translation performance is obtained for similar segment lengths. And in each case, a best performing translation is better than the baseline by atleast 0.6 BLEU points.

5.2. Effect of sentence length

Next we investigate the effect of original sentence length on segmentation. Table 3 shows segmentation-translation performance on BC data for different sentence length classes along with the optimal segment length (OSL) i.e, the segment length giving the highest BLEU score. The bin size was determined on the basis of the sentence length distribution. The baseline performance, which corresponds to manual segmentation, is also shown. Results show that for every sentence length class, BLEU scores improve with respect to the baseline. However, improvement in translation is more significant in case of longer sentences.

Table 3. Segmentation performance per sentence length class - BCAD05-D; *Mu - Average sentence length ; OSL - Optimal segment length in words; MS - Manual segmentation as baseline performance*

<8 words Mu: 5.98 MS: 8.02		8-15 words Mu: 11.92 MS: 7.23		16-30 words Mu: 19.83 MS: 7.89		>30 words Mu: 34.45 MS: 8.32	
OSL	BLEU	OSL	BLEU	OSL	BLEU	OSL	BLEU
5.9	**8.05**	10.6	**7.78**	16.9	7.70	25.8	8.35
5.8	8.04	10.0	7.43	15.2	7.68	22.6	8.55
5.4	7.78	8.6	7.27	12.3	7.87	14.6	8.65
5.3	7.75	7.3	7.58	8.7	**7.99**	9.0	8.66
4.9	7.49	6.3	6.90	6.9	7.63	7.1	**9.06**
4.6	7.35	5.4	6.98	5.6	6.45	5.8	8.54

Yet another observation is that irrespective of the sentence length class, the optimal segment length chosen is in the range of 8-10 words with the exception of the shorter sentence length class where the average length of the original sentences itself is 5.98 words. This suggests that the optimal segment length for translation depends on the translation system parameters rather than the length of input sentence. Table 4 shows the results for a similar analysis on the BN data. The overall trends are similar to those in BC data although optimal segment length is in the range of 10-12 words. We believe that this is due to the difference in the sentence structure between BN and BC, with BN having longer sentences than BC.

Table 4. Segmentation performance per sentence length class - BN data (RT04) ; *Mu - Average sentence length ; OSL - Optimal segment length in words; MS - Manual Segmentation as baseline performance*

<15 words Mu: 8.68 MS: 12.99		16-30 words Mu: 22.78 MS: 12.80		31-50 words Mu: 39.15 MS: 12.33		>50 words Mu: 64.54 MS: 12.00	
OSL	BLEU	OSL	BLEU	OSL	BLEU	OSL	BLEU
9.1	**12.99**	21.2	12.86	29.4	12.66	43.1	12.07
9.1	12.99	19.4	12.94	24.4	12.64	30.9	12.06
9.1	12.99	16.4	13.25	22.5	12.75	27.5	12.17
8.0	12.32	12.4	**13.53**	13.0	13.11	14.8	**12.55**
6.7	11.40	9.8	13.08	9.8	**13.63**	10.5	12.42
5.6	11.34	7.4	12.56	7.6	13.32	7.7	12.25
5.1	7.92	6.4	12.77	6.5	12.66	6.4	11.91
4.8	8.11	6.0	13.14	6.1	12.72	6.0	11.90

5.3. Segmenting ASR output in absence of known boundaries

In experiments reported so far, a known sentence was further segmented in order to achieve better translation performance. However, the option of segmenting within known sentence boundaries is not available in real world SLT systems. These systems take an audio stream (generally an entire show) which consists of sentences spoken by one or more speakers with no information on sentence boundaries. Prior to recognition, a speaker segmentation and clustering step is performed based on acoustic information. This step, which we shall hence refer to as *audio segmentation*, provides acoustic segments that are independently decoded by the ASR system. Therefore, in the absence of resegmentation of ASR output prior to translation, MT system has access to the ASR output segmented purely on the basic of acoustic information. This is suboptimal since specific sentence-like units and phrase boundaries are informed by lexical information in addition to the acoustics.

We used the sentence segmenter and information about Optimal Segment Length (OSL) derived in the previous section to resegment ASR output in the above mentioned scenario. Table 5. shows the results for BNAT05 and BCAD05-E with different segmentation strategies. Translation of show transcripts (WER=0) with human defined units (segmentation) gives best translation among the 4 cases. While the segmentation remains the same in going from row 1 to row 2, a WER of over 30% results in decrease in BLEU of over 15% for BC and 10% for BN. In row 3, ASR first-best with audio segmentation is translated. Since the WER is very close to that in case 2, the fall in BLEU of 8% for BN and 16% relative for BC can largely be attributed to poor segmentation. The fourth row corresponds to using the automatic segmenter to resegment the ASR-FB output stream (the OSL from prevous section is used). This gives an improvement of almost 0.8 BLEU points for BC and a slightly lesser 0.33 for BN. These results indicate that a segmentation scheme optimized for low WER (audio segmentation) does not necessarily hold

forth for translation. It is therefore necessary to resegment the ASR output stream taking into account lexical and acoustic cues in a way that maximizes translation performance.

Table 5. Translation performance in BLEU for different inputs and segmentation strategies on BNAT05 and BCAD05-E datasets ; ASR-FB: ASR first-best output; MS: Manual segmentation using human defined sentences ; Audioseg: Audio segmentation ; Auto-Seg: Automatic segmentation corresponding to OSL for development set ; WER - Word Error Rate in (%)

Input Type		Broadcast News		Broadcast Conversation	
Text	Segmentation	WER	BLEU	WER	BLEU
Transcripts	MS	0	16.53	0	14.11
ASR-FB	MS	30.30	14.66	31.46	11.93
ASR-FB	Audio	30.26	13.52	32.53	9.92
ASR-FB	Auto-Seg	30.26	13.85	32.53	10.67

In the previous section, resegmentation of ASR output within known sentence boundaries gave better translation performance compared to manual segmentation. However, similar gains are not seen when original sentence boundaries are not known as is observed by lower BLEU scores in row 4 compared to row 2 in Table 5. This is because in the absence of known sentence boundaries, automatic segmentation might result in segments across sentence boundaries. From the results in Section 5.1 and 5.3, one may conclude that the first step in automatic resegmentation of ASR output is to identify actual sentence boundaries accurately (recall as close to 100% as possible). In addition, resegmenting within these boundaries (even resulting in precision less than 100%) while maintaining, on average, an optimal segment length can give further improvement in translation performance.

6. CONCLUSION

In this paper, we show that a complete sentence is not an optimal unit for speech translation. This is because a sentence is generally composed of units that are coherent within themselves but are independent of each other as seen from a phrase-based MT system. Through experimental results on Arabic Broadcast Conversations and Broadcast News, we show when sentence boundaries are available, intra-sentence segmentation of ASR output can give upto 11% (on BC) and 6% (on BN) improvement in BLEU score compared to translating manually defined segments. We also show that audio segmentation optimized for low WER may not give the best translation performance and a sentence segmentation step prior to translation is necessary for better translation performance.

7. ACKNOWLEDGEMENTS

Thanks to Matthias Paulik for help with setting up the sentence segmentation tool.

8. REFERENCES

[1] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 1526–1540, September 2006

[2] Jing Huang and Geoffrey Zweig, "Maximum entropy model for punctuation annotation from speech," *In Proc. of ICLSP 2002*, pp. 917-920, 2002

[3] J. Xu, R. Zens, and H. Ney, "Sentence Segmentation Using IBM Word Alignment Model 1," *In Proceedings of the 10th Annual Conference of the European Association for Machine Translation (EAMT 2005)*, pp. 280-287, Budapest, May 2005.

[4] Takao Doi and Eiichiro Sumita, "Splitting Input for Machine Translation Using N-gram Language Model Together with Utterance Similarity," *Coling 2004*

[5] S.D. Kim, Byoung-Tak Zhang, Y. T. Kim, "Reducing Parsing Complexity by Intra-Sentence Segmentation Using Genetic Learning," *38th Annual Meeting of the Association for Computational Linguistics*, p.p 164-171, Hong Kong, 2000

[6] Alon Lavie, Donna Gates, Noah Coccaro and Lori S. Levin, "Input Segmentation of Spontaneous Speech in JANUS: A Speech-to-speech Translation System," *Workshop on Dialogue Processing in Spoken Language Systems*, pp. 86–99, 1996

[7] M. Ostendorf and D. Hillard, "Scoring structural mde: Towards more meaningful error rates," *EARS Rich Transcription Workshop*, 2004

[8] Mohamed Noamany, Thomas Schaaf, Tanja Schultz, "Advances in the CMU-InterACT Arabic Gale Transcription System," *Proceedings of the HLT/NAACL*, 2007

[9] Sebastian Stuker, Christian Fugen, Roger Hsiao, Shajith Ikbal, Qin Jin, Florian Kraft, Matthias Paulik, Martin Raab, Yik-Cheung Tam, and Matthias Wolfel, "The ISL TC-STAR Spring 2006 ASR Evaluation Systems," *In Proc. of the TC-STAR Workshop on Speech-to-Speech Translation*, Barcelona, 2006

[10] Matthias Eck, Ian Lane, Nguyen Bach, Sanjika Hewavitharana, Muntsin Kolss, Bing Zhao, Almut Silja Hildebrand, Stephan Vogel and Alex Waibel, "The UKA/CMU Statistical Machine Translation System for IWSLT 2006," *In Proc. of the IWSLT*, Kyoto, 2006

[11] Franz Och, "Minimum error rate training in statistical machine translation," *In Proc. of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp 160–167, Japan, July 2003

[12] K. Papineni and S. Roukos and T. Ward and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," *Technical Report RC22176, IBM Research Division, Thomas J. Watson Research Center*, 2001