OPTIMIZING SENTENCE SEGMENTATION FOR SPEECH 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 translates complete sentences, which are conveniently indicated by sentence boundary markers. However, translation of speech cannot rely on such boundary markers and therefore new methods are required that define an optimal unit for translation. In this paper we argue that translation performance can be improved by optimizing the translation segment length. Our experimental results show that choosing a segment length optimized for a particular MT system can obtain an improvement in BLEU score of up to 6% for Arabic broadcast news (BN) and 11% for broadcast conversation (BC) data, indicating that segment length optimization helps for planned as well as for conversational speech. We also observed significant degradation in translation performance with increasing word error rate (WER). Unfortunately, this degradation was not graceful. Since gains from segmentation are related to WER, the segmentation optimization becomes even more important. All these effects support our argument for a tighter coupling between ASR and MT systems.

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

1. INTRODUCTION

With significant growth in the performance of automatic speech recognition (ASR) over the past two decades, new problems in language technology are being pursued that use the output of an ASR system as the input for other applications. These applications include speech translation and summarization, reading tutors, dialogue systems and rich transcription tasks. 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 has significant bearing on the overall performance of the system.

Previous work in sentence segmentation has focused on spotting sentence boundaries as defined by humans and 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 prepared 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 computationally cumbersome, might not be optimum.

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 EBMT performance where longer sentence do not yield good translation. [5] proposed a technique to split sentences by matching sentences to those in corpus using editing distance criterion and show improvement in EBMT performance. However, no results on effects of recognition errors were reported. In [4], 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 more difficult domains such as translation of broadcast news.

The goal of this paper is to show that sentence segmentation has to be optimized taking into account the downstream process that will be applied. We investiage optimizing segmentation to improve translation accuracy and show this approach improves end-to-end performance. We study the extent and nature of degradation in translation with increasing Word Error Rate (WER) and how optimizing segmentation for translation can have compensating effect on this degradation.

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 the complete sentence and in the second, a segment boundary was marked at commas and periods.

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

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.

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

sults in improved translation. However, speech translation systems work on output of an ASR system where no commas or period information is provided and notion of punctuation for spoken language is unclear as evidenced in significant interannotator disagreement [7]. Moreover, the errors in the recognition output also contribute to degradation in translation performance.

3. SENTENCE SEGMENTER

To perform sentence segmentation on ASR output, we use 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. In addition to this, using acoustic/prosodic features such as pitch and energy did not yield significant improvement over LM probabilities and pause duration.

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 considered 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 can be used to control the average number of segments per sentence.

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

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

4. EXPERIMENTAL SETUP

4.1. System overview

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. 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 Optimization (MERO) [10] on the development sets. Separate optimizations were performed for BN and BC shows.

4.2. Evaluation sets

We investigate the effect of segmentation on two different sets of Broadcast news and Broadcast conversion shows. The BC data comprised 4 Al-Jazeera shows (dated 2005-02-18, 2005-02-16, 2005-03-01 and 2005-03-11) provided by the LDC. These shows are typically moderated panel discussions of 30 minutes each. 3 shows were chosen as the test set and remaining one (2005-02-18) as the development set. None of these shows overlap with training data used in ASR/SMT or the segmentation module. In all, the testset comprised of 359 sentences with an average sentence length of 16 words. In BC data, we noticed that there were a few cases where sentences were grammatically incomplete. This was either due to hesitation on the part of the speaker or due to interruptions from a panelist or the show moderator. We processed the data to completely remove all such sentences. Although in real broadcast conversation scenarios, speaker overlap is rather common and a challenging problem, we shall not deal with such situations in this paper. For experiments on BN data, we used 2 shows of 20 minutes each from the RT-04 evaluation set (ALJ-20031208, DUB-20021211). These consisted of 157 sentences with the average sentence length being 33 words.

5. EXPERIMENTS AND DISCUSSION

5.1. Effect of sentence segmentation on translation

First, manually segmented 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 testset, 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 for each of 3 BC shows and 2 BN shows. 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 thus result in poorer performance. For both the shows, the optimum translation performance is obtained for similar

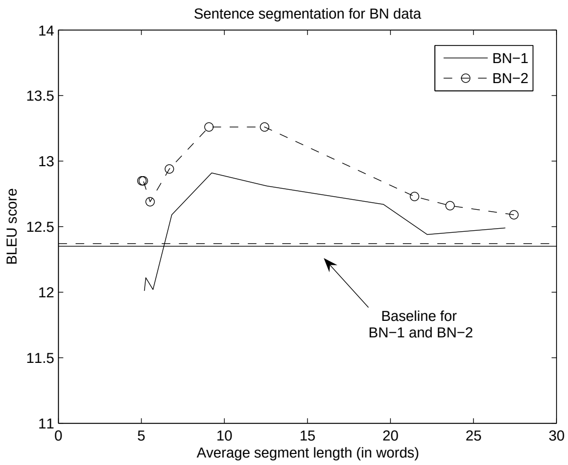

Fig. 1. Effect of segmentation on translation for Broadcast News; *Baseline performance (no segmentation) for BN-1 and BN-2 is marked*

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 investigated the effect of sentence length on segmentation for BC and BN data. Table 3 shows segmentation-translation performance on BC data for different sentence length classes along with the optimum segment length (OSL) i.e, the segment length giving highest BLEU scores. The bin size was determined on the basis of the sentence length distribution. The baseline performance, which corresponds to no 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 2. Segmentation performance per sentence length class - BC data; *Mu - Average sentence length ; OSL - Optimal segment length in words; NS - No Segmentation as baseline performance*

<8 words Mu: 5.98 NS: 8.02		8-15 words Mu: 11.92 NS: 7.23		16-30 words Mu: 19.83 NS: 7.89		>30 words Mu: 34.45 NS: 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 optimum segment length chosen is in the range of 8-10 words with the exception of the shorter sentence length class where the average length itself is 5.98 words. This tends to suggest that the optimal segment length for translation depends on the translation system parameters rather than the length of input sentence. Table 2 shows the results for a similar analysis for 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 structure of the sentence structure for BN and BC with BN.

Table 3. Segmentation performance per sentence length class - BN data; *Mu - Average sentence length ; OSL - Optimal segment length in words; NS - No Segmentation as baseline performance*

<15 words Mu: 8.68 NS: 12.99		16-30 words Mu: 22.78 NS: 12.80		31-50 words Mu: 39.15 NS: 12.33		>50 words Mu: 64.54 NS: 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. Effect of Word Error Rate

Since ASR is the first step in speech translation, recognition errors propagate through the MT system degrading translation performance. Thus a sentence with more errors is, on average, more likely to be inaccurately translated. So far however, the effect of WER translation performance has not been clearly established. We study this effect by comparing BLEU scores for sentences with different WERs as shown in Fig. 2. To avoid the interfering effects, no segmentation was performed, i.e. complete sentences were translated.

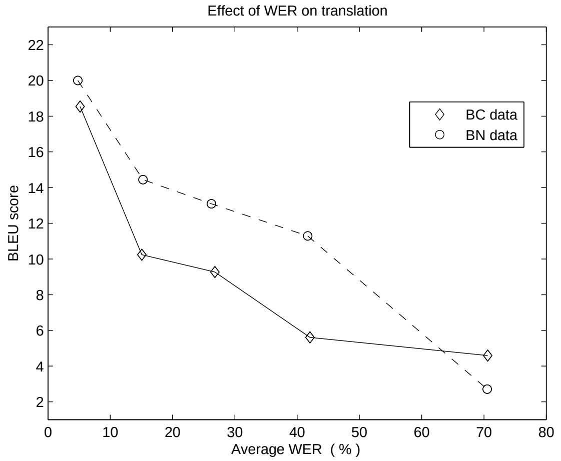

Fig. 2. Effect of Word Error Rate on speech translation

From Fig. 2, it is evident that there is significant degradation in translation performance with increasing WER, which is expected. However, the deterioration is not uniform. Initially, there is a steep fall in BLEU scores as the WER increases followed by a region where degradation is steadier. As the WER increases beyond 35%, translation performance drops steeply. This indicates that while improving WER should improve translation, the exact improvement itself depends on how well we are already doing in terms of WER.

We then investigated the effect of WER on end-to-end system performance, i.e. by performing automatic segmentation before translation. Table 4 and 5 show the translation performance and the corresponding optimum segment length for sentences in different WER classes in BN and BC data respectively. Also shown is the baseline performance, the distribution of WER and average WER for each class.

Table 4. Effect of WER on optimum segment length and BLEU for BN data. *OSL - Optimum segment length ; Bin size - percentage of total sentences ; OSL-seg. - segmentation with optimum segment length ; No seg. - No segmentation*

WER(Avg) range	Bin size (%)	OSL (words)	No seg. (BLEU)	OSL seg . (BLEU)
0-10 (4.80)	13.37	15.41	20.00	21.05
10-20 (15.26)	14.67	14.15	14.44	15.95
20-35 (26.24)	31.84	11.51	13.09	13.64
35-50 (41.69)	21.09	9.82	11.29	12.35
>50 (70.53)	19.10	5.03	2.71	3.75

We see that while segmentation improves translation performance with respect to the baseline for every WER class, higher gains are obtained for poorly recognized sentences. A related observation is that on average, shorter segments tend to be preferred as recognition errors increase. One of the possible reasons is that shorter segments tend to isolate ASR errors during translation and prevent error propagation across the sentence thus localizing the effect of the error. This effect is likely to be more pronounced when complex SMT decoders that permit longer-range phrase reordering are used. This points towards using ASR word confidence measures to guide translation, a topic for future study.

Table 5. Effect of WER on optimum segment length and BLEU for BC data. *OSL - Optimum segment length ; Bin size - percentage of total sentences ; OSL-seg. - segmentation with optimum segment length ; No seg. - No segmentation*

WER(Avg) range	Bin size (%)	OSL (words)	No seg. (BLEU)	OSL seg . (BLEU)
0-10 (5.15)	0.2	16.25	18.54	18.59
10-20 (15.06)	18.2	12.85	10.24	10.50
20-35 (26.80)	25.7	14.70	9.27	9.52
35-50 (42.06)	17.9	5.26	5.61	6.43
>50 (70.60)	31.3	4.49	4.59	5.27

6. CONCLUSION AND FUTURE WORK

In this paper, we show that a complete sentence is not an optimum 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 that segmenting ASR output optimizing for translation results in up to 11% and 6% improvement in BLEU score for BC and BN data respectively. We also study the degradation of MT performance with WER for BN and BC and hypothesize that segmentation can compensate ASR errors to a limited extent. In future work, we shall explore the use of ASR confidence measures to improve segmentation-translation performance thus moving towards tighter coupling of ASR and MT systems.

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] 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

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

[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 GALE 2006 evaluation system for Arabic broadcast news/conversation ASR,” *Submitted to ICASSP 2007*

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

[10] 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