SENTENCE SEGMENTATION AND PUNCTUATION RECOVERY FOR SPOKEN LANGUAGE TRANSLATION

Matthias Paulik[1,2]*, Sharath Rao*[1]*, Ian Lane*[1]*, Stephan Vogel*[1] *and Tanja Schultz*[1,2]

Interactive Systems Laboratories (interACT)
[1]Carnegie Mellon University (USA), [2]Universität Karlsruhe (Germany)
{paulik, skrao, ian.lane, stephan.vogel, tanja}@cs.cmu.edu

ABSTRACT

Sentence segmentation and punctuation recovery are critical components for effective spoken language translation (SLT). In this paper we describe our recent work on sentence segmentation and punctuation recovery for three different language pairs, namely for English-to-Spanish, Arabic-to-English and Chinese-to-English. We show that the proposed approach works equally well in these very different language pairs. Furthermore, we introduce two features computed from the translation beam-search lattice that indicate if phrasal and target language model context is jeopardized when segmenting at a given word boundary. These features enable us to introduce short intra-sentence segments without degrading translation performance.

Index Terms— Spoken Language Translation, Tight Coupling, Sentence Segmentation, Punctuation Recovery

1. INTRODUCTION

Spoken Language Translation (SLT) is traditionally separated into two independent components; Automatic Speech Recognition (ASR) and Machine Translation (MT). Within this pipeline, ASR provides an error-prone, audio segmented stream of non-punctuated words. In contrast, the majority of MT training data consists of non-speaking style bilingual text data with proper sentence segments and punctuation marks. To address the mismatch between ASR output and MT training data, it is possible to transform the ASR output towards the style of MT training data (punctuation recovery, disfluency removal, etc.). Another possibility is to transform MT training data towards the style of ASR output (punctuation removal, spelling out of numbers and dates, etc.) or to apply a mixture of both approaches. Sentence segmentation of ASR output and punctuation recovery prior to translation play a major role in this transformation process. Besides tackling the mismatch between ASR output and MT training data, sentence segmentation and punctuation recovery allows for better readability of the translation and may improve downstream language processing. Both of these benefits can be achieved by either inserting punctuation marks on the target side after translation or recovering punctuation on the source side.

In this paper we describe our sentence segmentation and punctuation recovery scheme applied to our GALE [1] 2007 Arabic-to-English (Ar→En) and Chinese-to-English (Ch→En) SLT systems. Beginning with our sentence segmentation / punctuation recovery system used during the TC-STAR [2] 2006 and 2007 evaluations, we first develop and test this scheme on our TC-STAR 2007 English-to-Spanish (En→Sp) SLT system and then transfer it to our GALE systems. Our approach introduces sentence segments marked with full stops as well as shorter, non-punctuated intra-sentence segments on the first-best ASR hypotheses. Further, it recovers commas on the target side during the translation process by applying modified phrase tables. In the second half of this paper, we introduce two novel features that indicate if phrasal context and target language model context is jeopardized when segmenting at a given word boundary. We analyze these features in the context of human style sentence segmentation and in the context of intra-sentence segmentation for machine translation.

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.

2. EXPERIMENTAL SETUP

2.1. Data

Our experiments on En→Sp used the official ASR hypotheses files that were provided during the TC-STAR 2006 and 2007 evaluations for the SLT task. These two data sets are referred to as *TC-Star06* and *TC-Star07*, respectively in this paper. ASR hypotheses were generated by rovering the ASR system outputs from the individual TC-STAR participating sides. Both evaluation sets had a WER (lowercase) of 6.9%. Translation references were case sensitive and included two references per source sentence.

For Ch→En and Ar→En two evaluation data sets, *dev-Test* and *eval06*, were used. These sets were extracted from the shadow data included in the ROSETTA team ASR output of the GALE 2007 evaluation. For Chinese, *dev-Test* consisted of 23K characters from the official 2007 development set and *eval06* consisted of 8K characters from the 2006 evaluation set. The CER was 10.5% and 17.1%, respectively. For Arabic *dev-Test* consisted of 8K running words from the 2007 development set and *eval06* consisted of 9k running words from the 2006 evaluation set. The WER was 12.1% and 21.7%, respectively. Translation references for the Ch→En and Ar→En GALE data were lowercase and consisted of only one reference per source sentence.

2.2. TC-STAR 2007 SMT System

The UKA TC-STAR 2007 En→Sp translation system for the SLT task was trained using the parallel European Parliamentary Speeches (EPPS) corpus provided within TC-STAR. Phrase tables were estimated via the GIZA++ toolkit [3] and University of Edinburgh's phrase model training scripts [4]. Part-of-speech based (POS) word reordering was applied to the source side prior to translation. During decoding a 4-gram language model (LM) built with the SRI LM

	TC-Star06	TC-Star07
baseline	38.70 / 46.21	36.57 / 45.58
fully mod.	39.23 / 45.84	38.29 / 45.39
mixed	**40.16 / 45.57**	**39.00 / 45.22**

Table 1. BLEU / TER scores for target side punctuation recovery via modified phrase tables on TC-STAR En→Sp.

toolkit and a 6-gram suffix array LM, both trained on the Spanish side of the EPPS corpus, were applied. A detailed description of a comparable UKA/CMU SMT system using the same architecture can be found in [5].

2.3. GALE 2007 Spoken Language Translation

For the GALE 2007 evaluation task, end-to-end SLT systems were developed for Ch→En and Ar→En language pairs. For transcription, multiple ASR systems were combined by applying cross-adaptation and confusion network combination. Phrase-based SMT systems were trained using an approach similar to that of the above mentioned TC-STAR system. For the Chinese and Arabic systems, no POS based reordering was applied. Instead, word reordering which assigns higher costs to longer distance reorderings was used. For the experiments reported in this paper, a reordering window of 4 and 2 was used for the Ar→En and Ch→En systems, respectively.

2.4. English Baseline Sentence Segmentation

The UKA/CMU sentence segmentation system [6], as it was used for the En→Sp SLT task within TC-STAR, inserts a full stop, comma or no punctuation mark at a given word boundary B_i based on the local language model context $w_{i-2}B_{i-1}w_iB_iw_{i+1}B_{i+1}w_{i+2}$. B_{i-1} is the boundary / punctuation mark type estimated in the previous step $i-1$ and for B_{i+1} all possible punctuation are being considered. An empirically estimated rule based on pause duration was used to constrain the insertion of punctuation marks. If the pause duration p is within $0.03s < p \leq 0.7s$ punctuation is estimated via the LM. A full stop is inserted for $p > 0.7s$. The LM was trained on the English EPPS corpus and segmentation was tuned to minimize WER on the TC-STAR 2005 development set.

3. COMMA RECOVERY VIA MODIFIED PHRASE TABLES

Punctuation recovery for speech translation can be performed either on the source side before translation or on the target side, during or after the translation process. In [7], the insertion of punctuation marks during the translation process is achieved by removing punctuation marks from the source side of the phrase table while retaining punctuation on the target side of the phrase table. It is pointed out that source punctuation should not be removed prior to word-alignment, since this may negatively affect alignment accuracy.
We compare this approach to the source side punctuation recovery performed in English baseline sentence segmentation (*baseline*), as described in Section 2.4. We investigate the target side recovery of full stops and commas (*fully modified*) as well as a mixed approach (*mixed*) in which we insert full stops on the source side before translation and commas on the target side during translation. In both cases, source sentence segmentation is defined by the baseline sentence segmenter, i.e. the input to MT differs only in the retained punctuation marks. The effectiveness of the two approaches compared to the baseline system is shown in Table 1. The mixed approach obtained the highest translation quality.

		dev-Test	eval06
Ch→En	baseline	8.31 / 80.45	8.72 / 77.82
	mixed	**8.77 / 79.58**	**10.09 / 76.64**
Ar→En	baseline	19.46 / 61.02	13.51/ 67.82
	mixed	**21.25 / 60.05**	**15.30 / 68.13**

Table 2. BLEU / TER scores for target side comma recovery via modified phrase table on GALE Ch→En and Ar→En.

	TC-Star06	TC-Star07
baseline segm.	40.16 / 45.57	39.00 / 45.22
mWER segm.	**41.40 / 44.11**	**41.25 / 42.82**
new automatic segm.	40.30 / 45.25	39.50 / 44.66

Table 3. Translation performance in BLEU / TER on En→Sp for tuning towards human style segmentation.

For the Ch→En and Ar→En GALE systems similar improvements in translation accuracy were gained when applying the mixed approach (see Table 2). For the GALE systems, source sentences were segmented according to the ASR audio segmentation with full stops inserted at the end of each segment. Subsequent experiments in this paper are conducted with accordingly modified phrase tables.

4. DECISION TREE BASED SENTENCE SEGMENTATION

4.1. Tuning towards Human Segmentation

In order to train a machine learning based sentence segmentation system, a set of 'ground truth' training examples is required. In this work we use punctuated ASR hypotheses during both training and testing. We automatically align the ASR hypotheses to the human provided transcripts by applying Universität Aachens multi word error rate (mWER) segmentation tool [8]. Thus, we achieve a human style segmentation of the ASR hypotheses towards which we tune the automatic sentence segmentation. Table 3 compares translation accuracy for the En→Sp system when translating mWER (human style) segmented ASR hypotheses, ASR hypotheses that were automatically segmented using the English baseline segmenter and hypotheses that were automatically segmented using the new decision tree based segmenter. Table 4 gives a similar comparison for the Ch→En and Ar→En GALE systems. The results show that it is desirable to tune the automatic sentence segmentation towards a human style sentence segmentation. In the following section we use F-Measure to compare the quality of a computed segmentation to the human style segmentation.

4.2. Feature Selection and Rule Creation

Our improved sentence segmentation architecture is based on a decision tree that uses multiple features computed for each boundary. We use J.R. Qinlan's C4.5 induction system [9] for decision tree training and rule extraction. We trained decision trees on different feature set combinations for the individual languages and selected the decision tree / feature set combination that yielded the highest F-Measure in regards to the human style segmentation. For English, we trained the decision tree on the *TC-Star06* ASR hypotheses and the final feature set combination consisted of word duration of the word preceding the current boundary, pause duration and LM probabilities for comma and full stop insertion (with the same local context as described in Section 2.4).

		dev-Test	eval06
Ch→En	audio segm.	8.77 / 79.58	10.09 / 76.64
	mWER segm.	**9.38 / 78.82**	**10.97 / 75.46**
	automatic segm.	8.92 / 79.58	10.73 / 75.97
Ar→En	audio segm.	21.25 / 60.05	15.30 / 68.13
	mWER segm.	**21.80 / 59.11**	**15.96 / 66.97**
	automatic segm.	21.40 / 59.67	15.53 / 67.71

Table 4. Translation performance in BLEU / TER on Ch→En and Ar→En for tuning towards human style segmentation.

	TC-Star06	TC-Star07
baseline segm.	54.79	52.48
new automatic segm.	65.97	62.14

Table 5. F-Measure of old and new automatic sentence segmentation for English.

For Chinese, we trained the decision tree on shadow data included in the ROSETTA team 2007 ASR dry-run. This data consisted of 6 shows from the 2006 development set and of the second half of the 2007 development set that was not included in *dev-Test*. For Arabic, we used 4 shows from the BNAD05 data set. For both languages, the final feature set combination consisted of pause and word duration as well as LM probabilities for full stop insertion. For Arabic, we also included a prosody based feature. Specifically, we encoded pitch information by combining pitch and delta pitch values in the vicinity of 700 milliseconds of the candidate boundary. We also included the signal power values in the same region as well as total signal power on either side of the boundary. As high dimensional features cause data sparsity problems and result in overfitting of the decision tree, we reduced the dimensionality by training a SVM based classifier on these features. We then used the scores of the SVM classifier as features within the decision tree. The same prosody based features were also considered for English and Chinese sentence segmentation. However, for these languages we did not observe any further improvements in terms of F-Measure. Tables 5 and 6 compare the F-Measures of the decision tree based sentence segmentation with the F-Measures of the respective baseline segmentation. BLEU and TER scores are listed in Table 3 and 4.

5. INCORPORATING PHRASAL AND TARGET LM CONTEXT DURING SENTENCE SEGMENTATION

Different source side sentence segmentations lead to different source phrase matches and different target side language model histories during translation. Possible word and phrase re-orderings during translation are also affected. Thus, translation quality is directly influenced by source side sentence segmentation. For a better integration of source sentence segmentation and phrase based MT, we propose to apply knowledge derived from the translation beam-search lattice. The motivation is not to break up source phrases that are valuable for MT and also to pay attention to the target LM context during sentence segmentation.

		dev-Test	eval06
Chinese	audio segm.	30.75	31.59
	automatic segm.	59.16	53.38
Arabic	audio segm.	33.89	37.50
	automatic segm.	40.97	43.41

Table 6. F-Measure of ASR audio segmentation and automatic sentence segmentation for Chinese and Arabic.

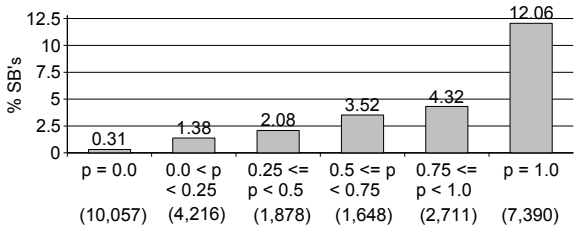

Fig. 1. Percentage of sentence boundaries found within different phrasal split-point probability ranges p. The overall number of boundaries (words) or each probability range is shown in brackets.

To compute a score indicating if phrasal context or target LM context is jeopardized when segmenting at a given word boundary, we apply a sliding window of 24 words with a step size of 6 words on the ASR output. For each step, we translate the 24 word sentence and compute two probabilities, $phrSP$ and $tbiSP$, for each of the 11 word boundaries between the innermost 12 words. These two probabilities are computed from the translation lattice used by our beam-search decoder. The edges in this lattice correspond to source words and phrases (together with their translation) and the nodes to the boundaries between these words and phrases. The phrasal split-point probability $phrSP$ for a given word boundary is computed as the number of paths going over its corresponding node divided by the number of paths visiting its node. We consider only the n-best pathes, i.e. the n-best translations. A phrasal split-point probability of one indicates that the word boundary is always seen between two source phrases in the n-best translations. Introducing a segment boundary at such a point should therefore not negatively affect possible phrase matches during translation. The target LM split-point probability $tbiSP$ is computed only for word boundaries with $phrSP > 0$ and is based on bi-gram probabilities. For all m word boundaries that are found to lie between two phrases, the target LM probability tbi of the bi-gram formed by the last word of the left source phrase and the first word of the right source phrase is computed. If the target LM does not include an according bi-gram, a bi-gram probability of 0 is assumed. $tbiSP$ is defined as: $tbiSP = 1 - (\sum^m tbi)/m$.

5.1. Experimental Evaluation

We analyzed the correlation of the phrasal split-point probability $phrSP$ with actual human sentence boundaries. We computed $phrSP$ for all word boundaries found in the human transcriptions of the *TC-Star06* set using the 100-best translations. We then selected six split-point probability ranges and computed the percentage of sentence boundaries found within these ranges. Figure 1 shows the result. While a high phrasal split-point probability does not necessarily predict a sentence boundary, a low phrasal split-point probability seems to be a strong indicator of a non-sentence boundary. However, augmenting our decision tree based sentence segmentation with $phrSP$ as an additional feature did not lead to any significant improvements. We repeated a similar experiment for the target LM split-point probability $tbiSP$. No clear correlation between $tbiSP$ and human sentence boundaries could be found.

6. INTRA-SENTENCE SEGMENTATION

Since the computed phrasal split-point probability indicates safe segmentation points in regards of MT phrase table knowledge, we examined the impact of introducing shorter intra-sentence segments at these split points on MT performance. The motivation for this was

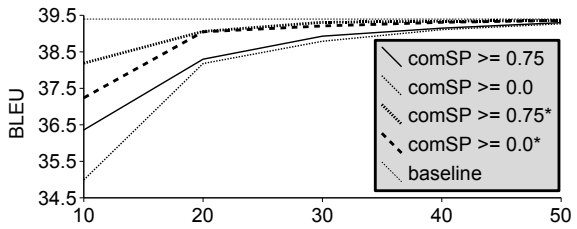

Fig. 2. BLEU scores for different segment lengths and different constraints on $comSP$. For the experiments marked with *, we retained the target LM context after each translation sentence.

on the one hand the demand of our confusion network translation for short source segments in order to keep memory consumption and run-time within reasonable boundaries. On the other hand, we followed the results presented in [10]. Here, it was shown that significant improvements in BLEU can be gained when introducing intra-sentence segmentation on top of given sentence boundaries.

To perform intra-sentence segmentation, we further segment the output generated by our decision tree based segmenter, by constraining the maximal allowed segment length. Each segment longer than this maximal allowed segment length is repeatedly split until the constrained is fulfilled. Only word boundaries with a combined phrasal and target LM split-point probability $comSP = a * phrSP + (1 - a) * tbiSP$ bigger or equal than x are considered and the final split point is selected with regards to pause duration and LM information. If no word boundary within the given segment fulfills the constraint $comSP \geq x$, the word boundary with the highest combined split-point probability is selected.

We computed BLEU and TER scores for different maximal segment lengths l with the combined split-point probability ($a = 0.5$) constrained to $comSP \geq 0.75$ and without any constrained on the combined split-point probability, i.e. $comSP \geq 0.0$. Results in BLEU for the En→Sp system on *TC-Star06* are shown in Figure 2. Due to decoder constraints, we only used a 4-gram SRI LM during translation, thus the overall translation performance is slightly lower than in Section 4. The results show that a significant advantage for computing $comSP$ is only given when creating short segments of less then 20 words per segment. However, for these small segment length we already observe a significant overall loss in translation performance due to the disrupted target LM context. For this reason, we changed our decoder to retain the LM state after decoding a sentence. By doing so, we are now able to constrain the maximal segment length to 30 words without any significant loss in BLEU (0.09 absolute) and with even a small improvement in TER (0.07 absolute). Furthermore, the degradation in MT performance for short segments of less than 20 words is now significantly smaller. While such short segments are not of interest for traditional SLT systems, they are necessary in the context of real-time simultaneous translation systems. As described in [11], such simultaneous MT systems require a continuous input of small and usefully translatable ASR hypotheses 'chunks' in order to achieve an acceptable MT performance while maintaining low latency.

7. CONCLUSION

We described our sentence segmentation and punctuation recovery scheme for spoken language translation. By applying modified phrase tables for implicit target side comma recovery during translation and by introducing a decision tree based sentence segmentation

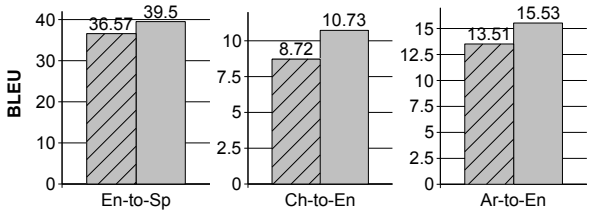

Fig. 3. Improvements in BLEU by applying our sentence segmentation and punctuation recovery scheme on our 2007 TC-STAR (eval07 data set) and GALE 2007 (eval06 data sets) systems.

for insertion of full stops on the source side, we significantly improved translation performance on three language pairs. Detailed results in BLEU are summarized in Figure 3. Furthermore, we investigated two novel features indicating if phrasal context and target language model context is jeopardized when segmenting at a given source word boundary. Incorporating these features enabled us to realize an intra-sentence segmentation that constrains the maximal segment length to 30 words without degrading translation performance and to achieve a graceful degradation in translation performance when translating short segments of less than 20 words. Such short intra-sentence segments are of interest for real-time simultaneous translation systems where translations have to be provided with a low latency.

8. REFERENCES

[1] GALE, "Global Autonomous Language Exploitation," http://www.darpa.mil/ipto/programs/gale/.

[2] TC-STAR, "Technology and Corpora for Speech to Speech Translation," http://www.tc-star.org.

[3] F.J. Och and H. Ney, "Improved Statistical Alignment Models," in *Proc. of ACL*, Hongkong, China, 2000.

[4] P. Koehn and C. Monz, "Manual and Automatic Evaluation of Machine Translation between European Languages," in *Proc. on the Workshop on SMT*, New York City, USA, 2006.

[5] M. Paulik, K. Rottmann, J. Niehues, S. Hildebrand, and S. Vogel, "The ISL Phrase-Based MT System for the 2007 ACL Workshop on SMT," in *Proc. of the ACL 2007 Second Workshop on SMT*, Prague, Czech Republic, June 2007.

[6] S. Stüker, C. Fügen, R. Hsiao, S. Ikbal, Q. Jin, F. Kraft, M. Paulik, M. Raab, Y. Tam, and M. Wölfel, "The ISL TC-STAR Spring 2006 ASR Evalutation Systems," in *Proc. of the TC-STAR Workshop on Speech-to-Speech Translation*, Barcelona, Spain, 2006.

[7] Y. Al-Onaizan and L. Mangu, "Arabic ASR and MT Integration For GALE," in *Proc. ICASSP*, Hawaii, USA, April 2007.

[8] E. Matusov, G. Leusch, O. Bender, and H. Ney, "Evaluating Machine Translation Output with Automatic Sentence Segmentation," in *Proc. of IWSLT*, Pittsburgh, PA, 2005.

[9] J.R. Quinlan, "The C4.5 Induction System - C4.5 Release 8," http://rulequest.com/Personal/.

[10] S. Rao, I. Lane, and T. Schultz, "Optimizing Sentence Segmentation For Spoken Language Translation," in *Proc. INTERSPEECH*, Antwerp, Belgium, August 2007.

[11] C. Fügen and M. Kolss, "The Influence of Utterance Chunking on Machine Translation Performance," in *Proc. INTERSPEECH*, Antwerp, Belgium, August 2007.