The Influence of Utterance Chunking on Machine Translation Performance

Christian Fügen, Muntsin Kolss

Interactive Systems Labs, Universität Karlsruhe (TH), Germany

fuegen@ira.uka.de, kolss@ira.uka.de

Abstract

Speech translation systems commonly couple automatic speech recognition (ASR) and machine translation (MT) components. Hereby the automatic segmentation of the ASR output for the subsequent MT is critical for the overall performance. In simultaneous translation systems, which require a continuous output with a low latency, chunking of the ASR output into translatable segments is even more critical. This paper addresses the question how utterance chunking influences machine translation performance in an empirical study. In addition, the machine translation performance is also set in relation to the segment length produced by the chunking strategy, which is important for simultaneous translation. Therefore, we compare different chunking/ segmentation strategies on speech recognition hypotheses as well as on reference transcripts.

Index Terms: machine translation, speech translation, simultaneous translation, segmentation, chunking

10.21437/Interspeech.2007-653

1. Introduction

In speech translation systems the combination of automatic speech recognition (ASR) and machine translation (MT) is not always straight forward, when optimal performance should be achieved. In addition to the errors committed by the speech recognition leading to additional errors in the machine translation, the ASR hypotheses have to be resegmented such that the performance of the MT does not suffer thereunder. Since almost all MT systems are trained on data split at sentence boundaries this is commonly done by resegmenting the hypotheses according to automatically detected sentence boundaries.

But automatic sentence boundary detection or punctuation annotation in general is, depending on the type of data, still very challenging. Punctuation annotation is usually done by combining lexical and prosodic features [1], whereas the combination is often done with the help of maximum entropy models [2] or CART-style decision trees [3]. Within TC-STAR [4] Lee et al. [5] proposed a system which inserts commas within a given ASR sentence by using ngram statistics for commas together with certain thresholds to improve the MT quality. [6] proposed another solution for inserting commas and periods into the ASR output by using a maximum entropy classifier using durational and language model features. As they observed on English a 98% correlation for periods and a 70% correlation for commas between the two and contiguous non-word sequences only such regions were considered.

In [7] different approaches for automatic sentence segmentation and punctuation prediction were compared with respect to MT performance. Punctuation prediction was either done with the help of a hidden ngram or by generating them implicitly during the translation process. For sentence segmentation an HMM-style search using hidden-events to represent segment boundaries was used, extended with an additional sentence length model. To obtain an optimal segmentation of a document a global search, restricted by the sentence length model has to be performed.

For simultaneous translation systems [8] chunking of ASR hypotheses into useful translatable segments is even more critical and difficult. Due to the resulting latency, a global optimization over several ASR hypotheses as suggested in [7] is impossible. Instead a maximum of 9-10 words resulting in a latency of about three seconds is desirable.

In this paper we extend the work of [9] in which the impact on translation performance of different text segmentation criteria was investigated. We address the questions on how chunking of ASR hypotheses as well as ASR reference transcripts into translatable segments, usually smaller then sentences, influence MT performance in an empirical study. Therefore, we compare different segmentation strategies on ASR hypotheses as well as on the reference transcripts. To measure the usefulness for simultaneous translation we set the MT performance in relation to the average segment length and its standard deviation.

2. Data and Systems

As test data we selected the 2006 Spanish-English TC-Star development data consisting of 3hrs (14 sessions) of non-native Spanish speech recorded at the European Parliament.We used ASR hypotheses as well as reference transcripts for the experiments, whereas the Spanish hypotheses were generated with a system trained within TC-STAR on Parliament Plenary Sessions [10]. The case-insensitive word error rate was 8.4%.

2.1. Statistical Machine Translation

The Spanish-English machine translation system [11] was trained using minimum error rate training on parallel European Parliamentary Speeches (EPPS) provided within TC-STAR and by Philipp Koehn [12].

For machine translation we used a phrase-to-phrase based statistical MT system. Various methods for phrase extraction have been proposed; in our system, phrase translation candidate pairs are extracted from the bilingual training corpus using the PESA method [13]. This method is suitable for open or large domain real-time translation systems, as phrase pairs of arbitrary length can be extracted from the bilingual corpus at decoding time, and does not require building a large static phrase table.

The decoder is a beam search decoder which allows for restricted word reordering. For our experiments, the following models were used: 1. a translation model, i.e. the word-to-word and phrase-to-phrase translations extracted from the bilingual corpus; 2. a trigram language model (LM); 3. a symmetric word reordering model, which penalizes longer-range reorderings by jump distance; 4. word and phrase count models which compensate the tendency of the LM to prefer shorter transla-

tions, and favor longer phrases over shorter ones, potentially improving fluency. Each of the model scores is multiplied by a scaling factor to give an overall score. The optimal set of model scaling factors is determined on a held-out set.

Decoding proceeds along the input segment, but allows reorderings of words and phrases by selecting, at each step, the next word or phrase to be translated from all words or phrases lying within a local window from the current position [14]. A window size of 4 was used in our experiments.

3. Experimental Results and Discussion

In this section we compare and discus the translation scores achieved by translating ASR reference transcripts as well as ASR hypotheses resegmented with different chunking strategies. Since punctuation annotation of ASR hypotheses is a research problem by itself and not the focus of this paper, all punctuation marks in the reference transcripts were removed for comparison reasons. However, the MT system was trained on complete sentences containing punctuation marks, since punctuation marks can provide useful alignment boundaries during the word alignment training.

Another problem is the influence of the LM on the translation quality of different chunking strategies. Since the LM is usually trained on sentences, chunking strategies producing segment boundaries with a high correlation to sentence boundaries are preferred and the influence and therefore the decrease in MT score might be higher the smaller the segments. But, depending on the chunking strategy a resegmentation of the LM training corpus in accordance to the chunking strategy is impossible. Therefore, in addition to the translation scores and segment length statistics we also give Precision and Recall by aligning the segment boundaries to punctuation marks in the ASR reference transcripts. A resegmentation of the LM training corpus was only done for the most appropriate chunking strategies and will be compared in Section 3.6.

3.1. Scoring MT with different Segmentations

The commonly used metrics for the automatic evaluation of machine translation output, such as the Bleu [15] and NIST [16] metrics, have originally been developed for translation of written text, where the input segment boundaries correspond to the reference sentence boundaries. This is not the case for translation of spoken language where the correct segmentation into sentence-like units is unknown and must be produced automatically by the system. In order to be able to use the established evaluation measures, the translation output of the automatically produced segments must be mapped to the reference translation segments in advance to the scoring procedure. This is done by using the method described in [17], which takes advantage of the edit distance algorithm to produce an optimal resegmentation of the hypotheses for scoring which is invariant to the segmentation used by the translation component.

The Bleu scores presented in this paper were obtained by using this method using two reference translations. Since the alignment was performed on a per session level the result is invariant in relation to the number of segments produced for a single session. The scoring was done case-insensitive without taking punctuation marks into account.

3.2. Baselines

Resegmenting ASR hypotheses at sentences boundaries for MT is the most common approach for speech translation systems. For this reason, the translation scores obtained by translating ASR hypotheses as well as reference transcripts split at sentence boundaries (*sent*) serve as one baseline for the following experiments. As can be seen in Table 1 we obtained a Bleu score of 36.6% by translating ASR reference transcripts and a score of 33.4% for ASR hypotheses, which clearly shows the influence of the ASR performance on MT quality. The average segment length was 30 words with a standard deviation of 22.

Another baseline is obtained by taking all punctuation marks as split points (*punct*). Thereby, the average segment length could be reduced to acceptable 9 words with almost no decrease in the translation score. The reason for that is, that punctuation marks are usually used to represent semantical boundaries. However, since the use of punctuation marks differ from language to language, they might be impractical as split points for other languages than Spanish. Furthermore, automatic punctuation annotation or even automatic semantic analysis is also impractical for simultaneous translation, because they are always erroneous and might require optimization over complete sentences [2, 3, 5, 7]. Therefore, in the following sections we analyzed how MT performance is affected by chunking strategies using other features and approaches requiring a smaller amount of context information for their decision.

3.3. Destroying the Semantic Context

In this section we analyzed, how MT performance is affected by destroying the semantic context of an utterance independently of the applicability for simultaneous translation. Following the experiments in [9] we simply cut the merged utterances of a single session every n word (*fixed*). The results are given in Table 1 for $n \in \{7, 11, 15\}$ and provide a lower bound for the chunking strategies that will be presented below. As expected, the decrease in segment size, i.e. the destruction of the semantic context affected the translation scores significantly. The translation results could be significantly improved by just cutting a sentence into two (*sent-0.5*) or four (*sent-0.25*) segments of equal size and not splitting across sentence boundaries. This clearly shows the dependency of the MT performance to the segmentation used for training the MT system.

3.4. Using Acoustic Features

Following the studies in [18] that pauses closely correspond to punctuation we used the information about non-speech regions in the ASR hypotheses for resegmentation. As non-speech regions we used recognized silences and non-human noises, whereas successive noises and silences were merged together. For the translation scores (*pause*) in Table 1 we used different non-speech duration thresholds (0.1, 0.2, and 0.3 seconds). As expected, the results are significantly better than those obtained with the chunking strategies in Section 3.3. The Precision and Recall values clearly validate the studies in [18] also for Spanish. While a threshold of 0.1 has the best correlation to punctuation marks, the MT score is the worst.

The standard deviations of the segment lengths achieved with the non-speech chunking strategies are still to large for the use in simultaneous translation. By splitting the ASR hypotheses at the longest non-speech interval within a region of a maximum number of words (*var15*, *var20*, *var25*, with chunks of maximal 15, 20, 25 words), the standard deviation could be significantly reduced without decreasing the translation quality when comparing fixed and variable non-speech chunking strategies having a similar average segment length.

For comparison reasons we evaluated also the performance

of our ASR segmentation (*asr*) which was developed in the context of the 2006 TC-Star evaluation [19]. Since it uses also non-speech regions satisfying some durational constraints as splitting points and performs a global optimization over the complete session, we expected that it should outperform the other non-speech based chunking strategies. A multi-layer perceptron was used for speech/ non-speech classification. While this chunking strategy is inapplicable for simultaneous translation, nevertheless the results are interesting. Compared to the baseline the degradation is less than one Bleu point without using any lexical/ semantic features. As can be seen as well, we measured a relatively high Precision of 71% and a Recall of 43%, which might be responsible for the good results.

Overall, chunking strategies using non-speech regions are simple and require no additional context information, but nonetheless achieving relatively good translation scores. While Precision and Recall show a good correlation between non-speech regions and punctuation marks only a slight correlation between Precision and Bleu score could be observed. This let us come to the conclusion that an optimal chunking strategy for MT does not necessarily has to have a high correlation with punctuation marks. Instead additional features have to be considered as well.

3.5. Using Other Features

In [20] prosodic as well as lexical information (punctuation marks, filled pauses and human noises) was used to automatically detect semantic boundaries. According to [20] a trigram language model was trained, whereas all punctuation marks in the training corpus were substituted by a boundary tag BD. But instead of doing a global optimization over the whole sentence the decision was made using local information only, i.e. (1) by setting the LM probabilities $Pr(w_{i-1}w_iBDw_{i+1}w_{i+2})$ and $Pr(w_{i-1}w_iw_{i+1}w_{i+2})$ in relation to each other and (2) by using additional thresholds for the non-speech gap in between w_i and w_{i+1}. As can be seen in Table 1 (*lm*) this chunking strategy outperforms all other strategies using acoustic features only. A Precision of 73% and a Recall of 54% was measured. Nevertheless a standard deviation of 10 from the average of 11 words was measured. Therefore, in a second experiment we relaxed the above mentioned thresholds when no split point was found after ten words (*lm-10*). Thereby, the standard deviation could almost be halved with only a minor decrease in Bleu score.

For the next experiment approached the problem in finding appropriate translatable segments for simultaneous translation from the other side. Instead of looking at the ASR hypotheses or references we looked at the PESA alignment information and at the reordering boundaries during the translation of the ASR hypotheses. Our hope was to find an optimal chunking with a small average segment length. Using the reordering boundary information provided by the MT system during a first translation of completely unsegmented ASR hypotheses, we split the ASR hypotheses at the reordering boundaries and retranslated them. As can be seen in Table 1 (*mt*) almost the same translation scores compared to the baseline could b reached, but the average segment length could be reduced to 17. For this chunking strategy we measured a Precision of 67% and a Recall of 33%.

3.6. Resegmenting the Language Model

As already mentioned above the LM is usually trained on sentences and is therefore negatively affecting the machine translation performance when using chunking strategies producing breaks independently from sentence boundaries. To achieve unaffected results the language model training data has to be resegmented according to the chunking strategy. Since the LM training data does not contain any acoustic information only lexical features like punctuation marks can be considered as split points. Therefore, we resegmented the LM training corpus by substituting each punctuation mark with a sentence boundary, i.e. newline and measured the MT performance for *punct*, *lm* and *lm-10*. Since the original LM training corpus contains punctuation marks we also measured the MT performance for *sent* with the help of a LM trained without punctuation marks on the original training data. Not surprisingly the MT scores are significantly better than compared to the results obtained with the original LM. But surprising is, that *punct* and even the automatic chunking strategies *lm* and *lm-10* outperform the original sentence based segmentation *sent*.

Chunking strategy	SegLength avg	sdev	Correlation Prc	Rcl	Bleu Ref	ASR
Baseline						
sent	30.1	22.1	98.5	26.5	36.59	33.41
punct	9.2	6.7	100.0	98.6	35.91	33.31
Length based						
fixed-7	7.0	0.2	22.8	26.5	30.13	27.50
fixed-11	11.0	0.7	22.6	16.8	32.07	29.53
fixed-15	15.0	0.6	23.8	13.0	33.64	30.66
sent-0.5	15.3	11.3	59.0	31.8	35.08	
sent-0.25	10.3	8.5	44.9	36.1	33.67	
Using acoustic features						
pause-0.1	8.2	5.7	59.3	58.3		31.86
pause-0.2	12.1	11.0	66.3	44.5		32.53
pause-0.3	17.0	19.6	71.5	34.0		32.62
pause-var15	7.5	3.1	59.4	61.4		31.34
pause-var20	9.8	4.3	64.6	53.0		31.87
pause-var25	11.8	5.3	68.3	46.9		32.36
asr	13.4	9.0	70.1	42.6		32.68
Using also other features						
lm	10.9	9.8	73.1	54.1		32.90
lm-10	8.7	5.6	67.3	62.4		32.36
mt	16.5	16.9	66.6	32.8		33.15
Using resegmented LM						
sent	30.1	22.1	98.5	26.5	37.56	34.28
punct	9.2	6.7	100.0	98.6	38.27	35.47
lm	10.9	9.8	73.1	54.1		34.93
lm-10	8.7	5.6	67.3	62.4		34.77

Table 1: Bleu scores obtained on ASR reference transcripts (Ref) as well as on ASR hypotheses (ASR) together with the average (avg) segment length and standard deviation (sdev). Precision (Prc) and Recall (Rcl) of segmentation boundaries to punctuation marks are given as well.

4. Conclusion

In this paper we have addressed the question on how utterance chunking influences machine translation performance in an empirical study by comparing different chunking strategies on ASR hypotheses as well as on ASR reference transcripts. As can be seen in Figure 1 sentence boundaries are a good criterion for utterance chunking, but are inapplicable for simultaneous translation because of the high average sentence length. Chunking strategies based on non-speech regions are simple and require no additional context information, but nonetheless achiev-

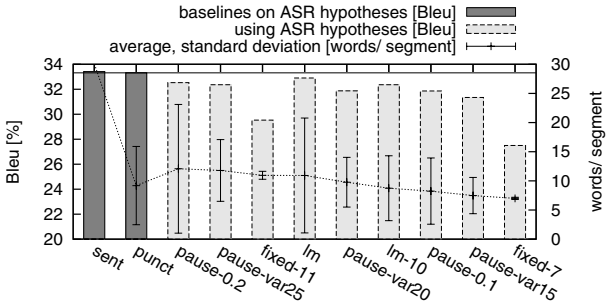

Figure 1: Results obtained by chunking ASR hypotheses sorted descending according to their average segment length. Bleu scores (left axe) are represented by boxes, the markers in the middle of the boxes give the average segment length (right axe) together with the standard deviation.

ing relatively good translation scores. Nevertheless, a more sophisticated approach using lexical features provided by a LM in addition outperforms all chunking strategies using acoustic features only. Furthermore, we have seen how important it is to resegment the LM training corpus for the MT system accordingly to the chunking strategy. Since acoustic/ prosodic information is usually not available for the LM training data, only lexical information can be used. Overall, the chunking strategy *lm-10* outperforms the manual sentence based segmentation and performs almost identical to the manual segmentation using punctuation marks *punct*. Furthermore, it has the desired average segment length of 9 and a small standard deviation of 6 and is therefore well suitable for simultaneous translation systems.

In the future, it might be worthwhile to approach the problem in finding an optimal chunking strategy from the other direction. Almost no decrease in translation quality could be achieved when using reordering boundaries taken from the a preliminary translation step as split points. But the problem how that segmentation could be imitated in advance to translation is still unsolved. Therefore, it might be also interesting for simultaneous translation systems to extend the MT decoder so that partial translations can be generated with only a short delay larger than the reordering window to reduce the latency.

5. Acknowledgments

We would like to thank Matthias Paulik for the ASR system and for his help in improving the MT system. This work has been funded by the *European Union* (EU) under the integrated project TC-STAR – Technology and Corpora for Speech to Speech Translation – (Grant number IST-506738).

6. References

[1] Y. Liu, *Structural Event Detection for Rich Transcription of Speech*, Ph.D. thesis, Purdue University, 2004.

[2] J. Huang and G. Zweig, "Maximum Entropy Model for Punctuation Annotation from Speech," in *Proc. ICSLP*, Denver, CO, USA, 2002.

[3] J.-H. Kim and P. C. Woodland, "The use of Prosody in a combined System for punctuation Generation and Speech Recognition," in *Proc. EUROSPEECH*, Aalborg, Denmark, 2001.

[4] TC-STAR, "Technology and corpora for speech to speech translation," 2004, http://www.tc-star.org.

[5] Y. Lee, Y. Al-Onaizan, K. Papineni, and S. Roukos, "IBM Spoken Language Translation System," in *TC-Star Speech to Speech Translation Workshop*, Barcelona, Spain, 2006.

[6] B. Ramabhadran, O. Siohan, L. Mangu, G. Zweig, M. Westphal, H. Schulz, and A. Soneiro, "The IBM 2006 Speech Transcription System for European Parliamentary Speeches," in *TC-Star Speech to Speech Translation Workshop*, Barcelona, Spain, 2006.

[7] E. Matusov, A. Mauser, and H. Ney, "Automatic Sentence Segmentation and Punctuation Prediction for Spoken Language Translation," in *Internat. Workshop on Spoken Language Translation*, Kyoto, Japan, 2006.

[8] C. Fügen, M. Kolss, M. Paulik, and A. Waibel, "Open Domain Speech Translation: From Seminars and Speeches to Lectures," in *TC-Star Speech to Speech Translation Workshop*, Barcelona, Spain, 2006.

[9] M. Cettolo and M. Federico, "Text Segmentation Criteria for Statistical Machine Translation," in *FinTAL – 5th International Conference on Natural Language Processing*, Turku, Finland, 2006, LNCS, pp. 664 – 673, Springer.

[10] S. Stüker, M. Paulik, M. Kolss, C. Fügen, and A. Waibel, "Speech Translation Enhanced ASR for European Parliament Speeches - On the Influence of ASR Performance on Speech Translation," in *Proc. ICASSP*, Honolulu, Hawaii, USA, 2007.

[11] M. Kolss, B. Zhao, S. Vogel, A. Venugopal, and Y. Zhang, "The ISL Statistical Machine Translation System for the TC-STAR Spring 2006 Evaluations," in *TC-Star Workshop on Speech-to-Speech Translation*, Barcelona, Spain, 2006.

[12] P. Koehn, "Europarl: A Multilingual Corpus for Evaluation of Machine Translation," 2003, http://people.csail.mit.edu/koehn/publications/europarl.

[13] S. Vogel, "PESA: Phrase Pair Extraction as Sentence Splitting," in *Machine Translation Summit 2005*, Thailand, 2005.

[14] S. Vogel, "SMT Decoder Dissected: Word Reordering," in *Int. Conf. on Natural Language Processing and Knowledge Engineering*, Beijing, China, 2003.

[15] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a Method for Automatic Evaluation of Machine Translation," Tech. Rep. RC22176 (W0109-022), IBM Research Division, T. J. Watson Research Center, 2002.

[16] NIST, "NIST MT evaluation kit version 11a," 2004, http://www.nist.gov/speech/tests/mt.

[17] E. Matusov, G. Leusch, O. Bender, and H. Ney, "Evaluating Machine Translation Output with Automatic Sentence Segmentation," in *Internat. Workshop on Spoken Language Translation*, Pittsburgh, USA, 2005.

[18] C. Chen, "Speech Recognition wiht Automatic Punctuation," in *Proc. EUROSPEECH*, Budapest, Hungary, 1999.

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

[20] M. Cettolo and D. Falavigna, "Automatic Detection of Semantic Boundaries based on Acoutic and Lexical Knowledge," in *Proc. ICSLP*, Sidney, Australia, 1998.