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 relatively simple chunking/ segmentation strategies on speech recognition hypotheses as well as on reference transcripts.

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

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 bigram/ trigram statistics for commas together with certain thresholds to improve the machine translation quality. [6] proposed another solution for inserting commas and periods into the ASR output by using a maximum entropy classifier. Durational and language model features were used for the classifier. 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 [8] 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 [9] chunking of ASR hypotheses into useful translatable segments is even more critical and difficult. Due to the resulting latency, a global optimization over the complete document or several ASR hypotheses as suggested in [7] is impossible. Instead a maximum of 9-10 words resulting in a latency of about 3 seconds is desirable.

In this paper we address the questions on how chunking of ASR hypotheses as well as ASR reference transcriptions into translatable segments, usually smaller then sentences, influence MT performance of a conventionally trained system, i.e. trained on complete sentences with punctuation marks, in an empirical study. Therefore, we compare different relatively simple 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 their standard deviation.

In Section 2 the topic of scoring the machine translation with different segmentations is addressed. Section 3 introduces the test data and the speech recognition and translation systems used for the experiments. In Section 4 the experimental results are presented and discussed. Finally, Section 5 concludes the paper.

2. Scoring Machine Translation with different Segmentations

The commonly used metrics for the automatic evaluation of machine translation output, such as the Bleu [10] and NIST [11] 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 before the scoring procedure. This is done using the method described in [12], which takes advantage of the edit distance algorithm to produce an optimal re-segmentation of the hypotheses for scoring which is invariant of the segmentation used by the translation component.

3. Data and Systems

As test data for our experiments we selected the 2006 Spanish-English TC-Star development data consisting of 3hrs of non-native Spanish speech recorded at the European Parliament divided into 14 sessions. 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 [13]. The case-insensitive word error rate was 8.4%.

3.1. Statistical Machine Translation

The Spanish-English machine translation system was trained on parallel European Parliamentary Speeches (EPPS) provided within TC-STAR and by Philipp Koehn [14]. For the language model an additional amount of 175M words of monolingual data collected from the web was used [15].

For machine translation we used a phrase-to-phrase based statistical machine translation 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 [16]. 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; 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 language model to prefer shorter translations, 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 [17]. A window size of 4 was used in our experiments.

4. 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. For comparison reasons all punctuation marks in the reference transcripts were removed. However, a conventionally trained MT system, i.e. trained on complete sentences containing punctuation marks, was used, because we were especially interested in how such a system is influenced by the different chunking strategies. We have also seen in the past, that such a system performs better than compared to one which is trained without any punctuation. The reason for that is, that punctuation marks are especially helpful during word alignment training, as they provide useful alignment boundaries. Overall this means, that the results obtained are worse than compared to those using punctuation marks and further could be slightly biased towards segmentation strategies which produce segments with a higher correlation towards punctuation boundaries. Therefore, in addition to the translation scores and segment length statistics we tried to measured also Precision and Recall by aligning the segment boundaries to sentence boundaries and commas in the ASR reference transcripts. For an end-to-end simultaneous translation system punctuation has to be inserted either in the source or in the target language only, for a better readability and understanding.

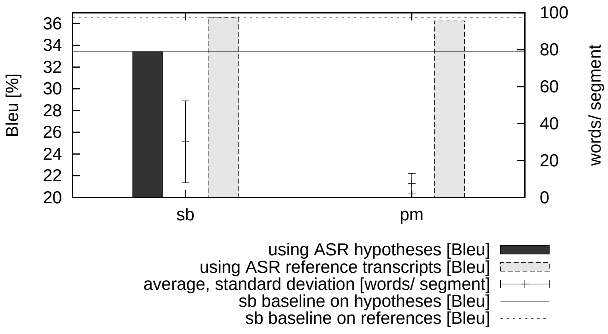

Figure 1: Baseline results achieved by splitting at sentence boundaries (sb) and punctuation marks in general (pm). Bleu scores (left axe) obtained by translating ASR reference transcripts (represented with gray boxes) as well as ASR hypotheses (represented with black boxes). The dashed line corresponds to the 'sb' baseline using reference transcripts, the solid line to the 'sb' baseline using ASR hypotheses. The vertical lines with markers in the middle of the boxes give the average segment length (right axe) together with the standard deviation.

The Bleu scores were obtained by using the method described in Section 2 using two reference translations. The scoring was done case-insensitive without taking punctuation marks into account.

4.1. Baselines

Resegmenting ASR hypotheses at sentences boundaries for machine translation 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 serve as one baseline for the following experiments. As can be seen in Figure 1 (sb) 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 (sentence) length was around 30 words with a standard deviation of 22.

Another baseline is obtained by translating ASR reference transcripts and taking also all other punctuation marks as additional split points. Thereby (Figure 1) the average segment length could be reduced from 30 to 8 with almost no decrease in the translation score. This is an unachievable and intransferable baseline, because the location of punctuation marks and therefore the segment lengths are language dependent and automatic punctuation annotation is always erroneous. Also automatic semantic analyses are very difficult. Therefore, in the following sections, we analyzed how MT performance is affected by chunking strategies using other features.

4.2. Destroying/ Extending the Semantic Context

In this section we analyzed, how MT performance is affected by destroying or extending the semantic context of an utterance independently of the applicability for simultaneous translation.

For this purpose, we first cut down the sentences into smaller chunks of equal length or merged several sentences to larger segments. As can be seen in Figure 2 while destroying the context by splitting a sentence four (s0.25) or two (s0.5)

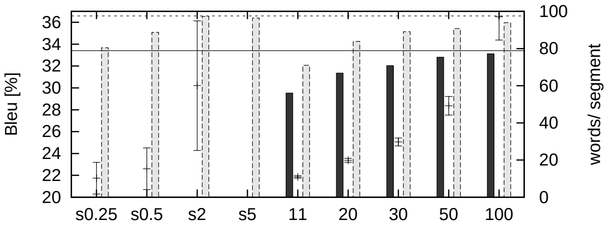

Figure 2: Results achieved by splitting (s0.25, s0.5) and merging sentences (s2, s5) and by splitting a complete session into chunks of fixed length (11, 20, 30, 50, 100). Legend given in Figure 1.

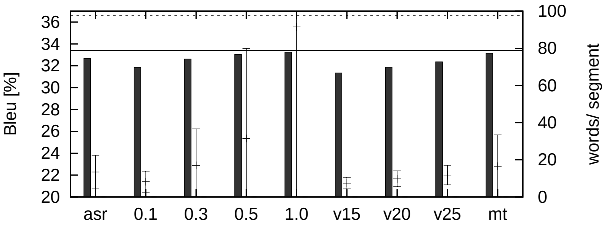

Figure 3: Results achieved by using the automatic ASR segmentation (asr), splitting at non-speech regions with different durational thresholds (0.1, 0.3, 0.5, 1.0), and by splitting at the longest non-speech interval within a region of a max. number of words (v15, v20, v25). In the last column, the results obtained by using information from an MT system are given (mt). Legend given in Figure 1.

times significantly decreases the MT scores, we can not make use of the extended context over sentence boundaries when merging two (s2) or five (s5) sentences together. This might be connected with the way of how the translation models were trained. Second we analyzed the additional influence of splitting across sentence boundaries on the translation score. Therefore, we merged all utterances of a single session together and cut them every n words. The results are given in Figure 2 for $n \in \{11, 20, 30, 50, 100\}$. As expected, the decrease in the segment size, i.e. the destruction of the semantic context affected the translation scores significantly when translating ASR hypotheses or reference transcripts. In comparison to the sentence merging and splitting and especially when comparing the translation scores across nearly equal sentence lengths the scores are worse, which means, that keeping at least the sentence boundaries seems to be important.

4.3. Using Additional Information

In the next set of experiments we analyzed how well other information extracted from the audio signal or ASR hypotheses is suitable for utterance chunking. In a first experiment we used the ASR segmentation which was developed in the context of the 2006 TC-Star evaluation [18]. The splitting was done at classified non-speech regions satisfying some durational constraints, while not throwing away any part of available speech. A multi-layer perceptron is used for speech/ non-speech classification. As this is done in multiple passes, it is inapplicable for simultaneous translation. But the result given in Figure 3 (asr) is interesting, because the degradation compared to the baseline is less than one Bleu point without doing any sentence boundary detection using e.g. linguistic or other features in addition. The reason might be due to the relatively high Precision of 71% when aligning the segmentation boundaries with the punctuation marks. The Recall was 43%.

Motivated by the previous results, 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 in Figure 3 we used different non-speech duration thresholds (0.1, 0.3, 0.5, and 1.0 seconds). As expected, the results are significantly better than those obtained with the chunking strategies in Section 4.2. For a threshold of 0.1 we observed a Precision of 60% and a Recall of 58%, for a threshold of 0.3 we observed 71% and 34% and for a threshold of 0.5 81% and 21%. While a threshold of 0.1 has the best correlation to punctuation marks, the MT score is the worst.

Since this chunking strategy is quite simple and requires no additional context information, but nonetheless achieving relatively good translation scores, it might be suitable even for simultaneous speech translation systems. The only problem is the large standard deviation. By splitting the ASR hypotheses at the longest non-speech interval within a region of a maximum number of words, the standard deviation could be significantly reduced (v15, v20, v25, with chunks of maximal 15, 20, 25 words) without decreasing the translation quality compared to the results achieved with a non-speech duration threshold of 0.3. When aligning the segmentation boundaries to punctuation marks for the experiments, we obtained for v15 a Precision of 59% and a Recall of 61%, for v20 64% and 53%, and for v25 68% and 47%.

Overall, the Precision and Recall values for the alignment of segmentation boundaries with punctuation reflect the well-known correlation between non-speech regions and punctuation marks. But in comparison with the MT, there seems to be no direct correlation between the Bleu scores and Precision and Recall. 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 semantic boundaries have to be found, which are also known to be correlated with non-speech regions or other prosodic features [19]. Besides non-speech regions also hesitations should be taken into account. For simultaneous speech translation an equally good performing chunking strategy comparable to the baseline with a small average segment length could not yet be found.

4.4. Using Machine Translation Information

For this experiment we would like to approach 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.

In a first experiment we used that information to split the ASR hypotheses at PESA alignment boundaries. This was not successful, because the average phrase length was about 2 words only, which means that the necessary context information for a good translation was almost lost and word reordering no more possible. The short phrase lengths can be explained by the domain and speaking style mismatch between the training and test data. While the European Parliamentary Speeches are mostly planned speeches and therefore more fluent, the test data is more spontaneous.

In a second experiment we used the reordering boundaries instead, i.e. we split the ASR hypotheses so that the reordering was not affected. As shown in Figure 3 (mt) nearly the same translation scores could be reached compared to the baseline, 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%.

5. Conclusion

In this paper we have addressed the question on how utterance chunking influences machine translation performance in an empirical study by comparing different relatively simple chunking strategies on ASR hypotheses as well as on ASR reference transcripts. We have seen that sentence boundaries are a good criterion for utterance chunking, but are inapplicable for simultaneous translation because of the high average sentence length. At least for Spanish, punctuation marks in general seems to be more suitable for utterance chunking for simultaneous translation, but that raise the question of the transferability to other languages. Furthermore, punctuation marks are difficult to detect automatically. In contrast thereto, non-speech regions, which do not have to be correlated with punctuation marks seem to be a good indicator for a split. When limited to a maximum length, within which a split has to be occur, the decrease in translation quality is in our opinion tolerable and therefore suitable for simultaneous translation.

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. The average segment length could be significantly improved compared to the baseline, albeit still inappropriate for simultaneous translation. Furthermore, the problem how that segmentation could be imitated in advance to translation is still unsolved.

6. 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).

7. 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] A. Stolcke, E. Shriberg, R. Bates, M. Ostendorf, D. Hakkani, M. Plauche, G. Tr, and Y. Lu, “Automatic detection of sentence boundaries and disfluencies based on recognized words,” in *Proc. ICSLP*, Sidney, Australia, 1998.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

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

[15] M. Kolss, B. Zhao, S. Vogel, A. Venugopal, and Y. Zhan g, “The ISL Statistical Machine Translation System for t he TC-STAR Spring 2006 Evaluations,” in *TC-Star Workshop on Speech-to-Speech Translation*, Barcelona, Spain, 2006.

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

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

[18] 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.

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