Incremental Segmentation and Decoding Strategies for Simultaneous Translation

Mahsa Yarmohammadi[†], **Vivek K. Rangarajan Sridhar**[°], **Srinivas Bangalore**[°], **Baskaran Sankaran**[‡]
[†]Center for Spoken Language Understanding, Oregon Health & Science University
[°]AT&T Labs - Research
[‡]School of Computing Science, Simon Fraser University
{yarmoham}@ohsu.edu, {vkumar,srini}@research.att.com, {baskaran}@cs.sfu.ca

Abstract

Simultaneous translation is the challenging task of listening to source language speech, and at the same time, producing target language speech. Human interpreters achieve this task routinely and effortlessly, using different strategies in order to minimize the latency in producing target language. Toward modeling the human interpretation process, we propose a novel input segmentation method using the phrase alignment structure of the language pair. We compare and contrast three incremental decoding and two different input segmentation strategies, including our proposed method, for simultaneous translation. We present accuracy and latency tradeoffs for each of the decoding strategies when applied to audio lectures from the TED collection.

1 Introduction

In simultaneous speech translation, it is important to keep the delay between a source language chunk and its corresponding target language chunk (referred to as *ear-voice span*) minimal in order to continually engage the listeners. Simultaneous human interpreters are able to generate target speech incrementally with very low ear-voice span by using a variety of strategies (Chernov, 2004) such as anticipation, cognitive and linguistic inference, paraphrasing, etc. However, current methodologies for simultaneous translation are far from being able to exploit or model such complex phenomena. Quite often, models trained for consecutive translation are repurposed for incremental translation.

One of the first attempts at incremental *text* translation was presented by Furuse and Iida (1996) using a transfer-based MT approach and more recently by Sankaran et al. (2010) using a phrase-based approach. On the other hand, incremental *speech* translation has been addressed in simultaneous translation of lectures and speeches (Hamon et al., 2009; Fügen et al., 2007). Some previous work (Cettolo and Federico, 2006; Rao et al., 2007; Matusov et al., 2007) addressed source text (reference or ASR hypothesis) segmentation strategies in speech translation. Constraining the search process during decoding to be monotonic (Tillmann and Ney, 2000) is one way of reducing latency and promoting incrementality. However, finding the optimal segmentation of the complete source string using dynamic programming is still slow.

By shifting the focus of the task to appropriate segmentation of incoming text, consecutive translation models have been used with good success to simulate incremental translation, such as incremental speech-to-speech translation (Bangalore et al., 2012) which focuses on translating the partial hypotheses generated based on the silences detected by a speech recognizer. However, studies on human interpreters show that in only a few cases the interpreters encode the chunks of speech as uttered in the source: the mean proportion of silence-based chunking by interpreters is 6.6% when the source is English, 10% when it is French, and 17.1% when it is German (Pöchhacker, 2002). As an alternative to silence-based segmentation, in this work, we propose a novel approach for segmenting the incoming text that exploits the alignment structure between words (phrases) across a language pair. We compare the two segmentation methods in three different decoding strategies. We perform our investigation within an English-French phrase-based speech translation system trained and tested on TED talks released as part of the IWSLT evaluation (Federico et al., 2011).

2 Non-incremental and Incremental Translation

The objective in machine translation is to translate a source sentence $\mathbf{f} = f_1^J = f_1, \cdots, f_J$ into target sentence $\mathbf{e} = e_1^I = e_1, \cdots, e_I$. Given the in-

International Joint Conference on Natural Language Processing, pages 1032–1036,
Nagoya, Japan, 14-18 October 2013.

put sentence $\mathbf{f}$, we choose the sentence with highest probability among all possible target sentences. Since, it is intractable to estimate the conditional probability distribution $\Pr(\mathbf{e}|\mathbf{f})$ over sentences, we simplify the problem as mapping between sentential sub-units (words or phrases) and represent the correspondence across these units using an alignment structure, $\mathbf{a} = a_1^J = a_1, \cdots, a_J$.

$$\hat{\mathbf{e}}(\mathbf{f}) = \arg\max_{\mathbf{e}} \left\{ \sum_{\mathbf{a}} \Pr(\mathbf{e}, \mathbf{a}|\mathbf{f}) \right\} \quad (1)$$

In an incremental translation framework, we do not observe the entire string $\mathbf{f}$. Instead, we observe segments of the string. A sentence pair (f_1^J, e_1^I) can be segmented into K phrase pairs $\mathbf{s} = s_1^K = s_1, \cdots, s_K$,

$$s_k = (i_k; b_k, j_k) \ \ \forall \ \ k = 1, \cdots, K \quad (2)$$

where i_k is the end position of the word in target phrase k and (b_k, j_k) represent the start and end positions of the source phrase aligned with the target phrase k. To achieve the highest *monotonicity* in incremental translation, we may restrict the decoding problem to strictly generate *monotonic phrases* by satisfying the constraint, $b_k = j_{k-1}+1$ $\forall \ \ k = 1, \cdots, K$. We also constrain the source and target phrases to be ordered monotonically, meaning that if a source phrase at position j is translated to a target phrase at position i , then a source phrase at position $j' > j$ will be translated to a target phrase at position $i' > i$. We call such phrase pairs to be a *monotonic phrase alignment* for a sentence pair. Figure 1 shows an example of a word alignment matrix, all possible phrase pairs, and all possible monotonic phrase alignments (4 alignments) for the parallel sentences $\mathbf{e}$-$\mathbf{f}$, shown with different line styles. For instance, the monotonic phrase alignment shown with dark lines has three phrase pairs $s_1 = (0; 0, 0)$, $s_2 = (3; 1, 3)$, $s_3 = (4; 4, 5)$. Grey dotted-line phrases are not monotonic. In Section 3.2 we present a source sentence segmentation approach that makes use of the monotonic phrase alignments information.

3 Segmentation of ASR output for MT

In this section, we describe two alternative methods to split the input sentence into partial segments for incremental translation. Since the ASR component is not the main focus of our study, we do not explain the ASR system we used in detail. Our ASR system uses context-dependent HMMs

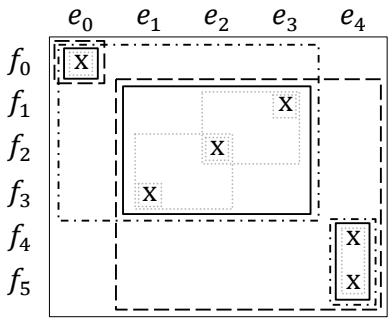

Figure 1: Word alignment matrix for two parallel sentences and their monotonic phrase alignments.

with Vocal Tract Length Normalization (VTLN) to build its acoustic model from 1119 talks we harvested from the TED website. We used the AT&T FSM toolkit (Mohri et al., 1997) to train a trigram language model for English from the permitted data in IWSLT 2011 evaluation. We reached 78.8% and 77.4% ASR word accuracies on the IWSLT dev2010 and tst2010 sets respectively.

3.1 Silence-based Segmentation

The output of automatic speech recognition includes silence information that is typically discarded before passing the source string into the machine translation component. We use any silence, irrespective of the frame length, as a segmentation marker. The average length of a segment using this strategy is 4.28±3.28 words.

3.2 Monotonic Phrase-Based Segmentation

In this section, we present an approach to split the source sentence into segments that can be monotonically translated to the target language. To prepare the training data for our segmentation model, we extracted monotonic phrase alignments from the set of all possible phrase alignments of a sentence pair in the word alignment matrix produced by GIZA++ using dynamic programming. We used 90% of the total parallel sentences and their extracted monotonic phrase alignments as training set, and reserved the rest 10% as development set. To get more meaningful alignments, we restricted those to the alignments of length at least 4.

Having the above training data, we trained a binary classifier, which was applied independently at each word in the sentence, to decide whether that word is a segment boundary or not. We used a discriminative log-linear model to train the classifier and we used the perceptron algorithm (Collins, 2002) to train the model parameters. Fisher and

Roark (2007), successfully used a discriminative log-linear model using the perceptron algorithm for automatic discourse segmentation task.

The task is to learn a mapping from inputs $x \in X$ to outputs $y \in Y$, where X is the set of sentences and Y is the set of possible monotonic alignments of the sentences. Given a set of training examples (x_i, y_i), a function $\mathbf{GEN}(\mathbf{x})$ that enumerates a set of possible monotonic alignments of x, $\bar{\alpha} \in \mathbf{R}^d$ a parameter vector, and representation Φ that maps each $(x, y) \in X \times Y$ to a feature vector $\Phi(x, y)$, there is a mapping from an input x to an output $F(x)$ defined by the formula:

$$F(x) = \underset{y \in \mathbf{GEN}(x)}{\arg\max}\ \Phi(x, y) \cdot \bar{\alpha} \qquad (3)$$

The model learns the parameter values $\bar{\alpha}$ during the training, and the decoding algorithm searches for the y that maximizes 3. The feature vector $\Phi(x, y)$ represents arbitrary features of the alignments. In our study, the feature set contains word, position of the word in the sentence, and segment length. For example, one feature might be (word='cat', position=8, seg_length=3, seg_boundary = true), which returns 1 if the current word is 'cat', it is the 8th word in the sentence, it is the 3rd word in the segment, and it is marked as a segment boundary, and returns 0 otherwise.

We evaluated our segmentation model with precision, recall and F1-score, defined in Eq. 4. Suppose a sentence of length n has m segment boundaries in the gold standard and k segment boundaries in the system output. Assume t out of k guessed boundaries are correct. Since we might have multiple valid segmentations for a sentence in our training data, we chose the gold standard to be the valid segmentation which has the minimum Levenshtein edit distance with the system output.

$$P = \frac{t}{k}, R = \frac{t}{m}, F1 = \frac{2PR}{P+R} = \frac{2t}{k+m} \qquad (4)$$

We achieved $P = 70.51\%$, $R = 91.52\%$, and $F1 = 75.89\%$ on the development set. The average length of a segment using this strategy is 6.56±4.73 words.

4 Decoding Strategies

We used three different decoding strategies for translating the ASR outputs. We tried each of these three techniques for incremental as well as regular (non-incremental) translation.

First, we used the Moses toolkit (Koehn et al., 2007) for statistical machine translation. Minimum error rate training (MERT) was performed on the development set (dev2010) to optimize the feature weights of the log-linear model used in translation. During decoding, the unknown words were preserved in the hypotheses. The parallel text for building the English-French translation model – around 6.3 million parallel sentences – was obtained from several corpora: Europarl (Koehn, 2005), jrc-acquis corpus (Steinberger et al., 2006), Opensubtitle corpus (Tiedemann and Lars Nygaard, 2004), WMT11 Gigaword (Callison-Burch et al., 2011), WMT11 News (Callison-Burch et al., 2011), and Web crawling (Rangarajan Sridhar et al., 2011) as well as human translation of proprietary data.

Second, we used a finite-state implementation of translation without reordering. We represent the phrase translation table as a weighted finite state transducer (FST) and the language model as a finite-state acceptor. The weight on the arcs of the FST is the dot product of the MERT weights with the translation scores. Our FST-based translation is the equivalent of phrase-based translation in Moses without reordering.

In addition to Moses and FST decoders, we used the incremental beam search decoder introduced by Sankaran et al. (2010) for translating in regular and incremental modes. This decoder modifies the beam-search decoding algorithm for phrase-based MT aiming at efficient computation of future costs and avoiding search errors. In Section 6 we show the results of translating our data using these three decoding strategies, referred to as Moses, FST and IncBeam decoders.

5 Data

In this work, we focus on the speech translation of TED talks. Over the past couple of years, the International Workshop on Spoken Language Translation (IWSLT) has been conducting the evaluation of speech translation on TED talks for English-French. We leverage the IWSLT TED campaign by using identical development (dev2010) and test data (tst2010).

6 Experiments and Results

We compare the results in terms of accuracy of translation and latency of generating partial outputs. We translated and evaluated each of 11 test

sets independently and we report the average values. In incremental mode, we ran Moses with *continue-partial-translation* option which enables chunk translation to be conditioned on history. In contrast, FST performs a chunk-wise translation which is independent of history.

		Moses	FST	IncBeam
Regular	ASR	18.67	18.11	17.73
	Transcript	22.66	22.11	21.32
Incr. silence seg.	ASR	17.41	16.88	17.33
Incr. monotone seg.	ASR	17.64	17.09	17.40

a) Reference translation has punctuations

		Moses	FST	IncBeam
Regular	ASR	23.04	22.58	22.00
	Transcript	28.38	27.75	26.63
Incr. silence seg.	ASR	21.66	21.12	21.38
Incr. monotone seg.	ASR	21.69	21.26	21.48

b) Reference translation has no punctuations

Table 1: Accuracy (BLEU) of English-French MT models on reference transcripts and ASR outputs

Table 1 shows translation accuracies in terms of BLEU scores. We consider the regular decoding as the baseline. Since we know the entire source input in advance, our baseline, obviously, has the highest accuracy but also the highest latency. For the baseline, we translated the ASR output and the reference transcript of the utterance. As shown in the "Regular" row, the accuracy on the ASR output drops by around 4% compared to that on the reference text. Since ASR outputs and the training data for our translation model do not contain punctuations, we also measured the accuracy against the references with removed punctuations.

Incremental translation of monotone-based segments gets a slightly higher accuracy than the silence-based segments for all the three decoders. In both regular and incremental decoding settings, the BLEU scores of Moses are higher than other two decoders. The FST decoder is better than the IncBeam decoder in regular setting; on the other hand the performance of the IncBeam decoder is better than the FST decoder and comparable to Moses in the two incremental settings. Both Moses and IncBeam decoders use reordering knowledge as well as history of translation in the incremental decoding settings, whereas the FST decoder lacks the latter.

In Table 2, we present the average speed of translating ASR output chunks. For each sentence the speed is calculated as the total time taken to translate the chunks divided by the number of chunks of the sentence. The speed reported in the table is then calculated by taking the average of speeds of all sentences in the test set. This measurement provides a good indication of latency in real-time translation. We note that we do not compare the delay of the decoders with each other due to differences in implementation and invoking the decoders, instead we compare the delays of each decoder by itself in three modes of translation.

	Moses	FST	IncBeam
Regular	2.35	2.06	17.68
Incr. silence seg.	0.68	1.75	6.43
Incr. monotone seg.	0.87	1.59	8.60

Table 2: Speed of generating target chunks (sec)

Comparing the accuracy values in Table 1 and latency values in Table 2 shows that in incremental decoding using the Moses and IncBeam decoders, we get some gain in accuracy but we loose some speed in monotone-based model compared to the silence-based model.

The interesting achievement is in incremental translation of monotone-based segments using the FST decoder. In this condition, we not only achieve an improvement in accuracy, but we also get a reduction in latency compared to the translation of silence-based segments. When translating each chunk independently, a meaningful segmentation of the input toward increasing the monotonicity yields a better performance in simultaneous translation than a silence-based segmentation.

7 Conclusions

In this paper we introduced a novel incoming text segmentation approach aiming at increasing the monotonicity of simultaneous translation. Using our proposed framework, we could achieve a point in segmenting and decoding the ASR output which enables simultaneous speech translation with a good accuracy/latency trade-off, even without relying on the history of translation. For future work we plan to improve our monotone-based segmentation model by using richer feature sets which for example include syntactic knowledge of the language. We are also interested in exploring our techniques on translating the languages with different word orders such as English/Japanese.

Acknowledgments

We would like to thank Brian Roark for his valuable discussions.

References

S. Bangalore, V. K. Rangarajan Sridhar, P. Kolan, L. Golipour, and A. Jimenez. 2012. Real-time incremental speech-to-speech translation of dialogs. In *Proceedings of NAACL:HLT*, June.

C. Callison-Burch, P. Koehn, C. Monz, and O. Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64, Edinburgh, Scotland, July. Association for Computational Linguistics.

M. Cettolo and M. Federico. 2006. Text segmentation criteria for statistical machine translation. In *Proceedings of the 5th international conference on Advances in Natural Language Processing*.

G. V. Chernov. 2004. *Inference and anticipation in simultaneous interpreting*. John Benjamins.

M. Collins. 2002. Discriminative training methods for hidden markov models: theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, EMNLP '02, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

M. Federico, L. Bentivogli, M. Paul, and S. Stüker. 2011. Overview of the IWSLT 2011 evaluation campaign. In *Proceedings of IWSLT*.

S. Fisher and B. Roark. 2007. The utility of parse-derived features for automatic discourse segmentation. In *In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 488–495.

C. Fügen, A. Waibel, and M. Kolss. 2007. Simultaneous translation of lectures and speeches. *Machine Translation*, 21:209–252.

O. Furuse and H. Iida. 1996. Incremental translation utilizing constituent boundary patterns. In *In Proc. of Coling '96*, pages 412–417.

O. Hamon, C. Fügen, D. Mostefa, V. Arranz, M. Kolss, A. Waibel, and K. Choukri. 2009. End-to-end evaluation in simultaneous translation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, March.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. J. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL*.

P. Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*.

E. Matusov, D. Hillard, M. Magimai-Doss, D. Hakkani-Tür, M. Ostendorf, and H. Ney. 2007. Improving speech translation with automatic boundary prediction. In *Proceedings of Interspeech*.

M. Mohri, F. Pereira, and M. Riley. 1997. Att general-purpose finite-state machine software tools, http://www.research.att.com/sw/tools/fsm/.

F. Pöchhacker. 2002. *The Interpreting Studies Reader*. Routledge (Taylor and Francis), New York.

V. K. Rangarajan Sridhar, L. Barbosa, and Bangalore. S. 2011. A scalable approach to building a parallel corpus from the web. In *INTERSPEECH*, pages 2113–2116.

S. Rao, I. Lane, and T. Schultz. 2007. Optimizing sentence segmentation for spoken language translation. In *Proceedings of Interspeech*.

B. Sankaran, A. Grewal, and A. Sarkar. 2010. Incremental decoding for phrase-based statistical machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, WMT '10, pages 216–223, Stroudsburg, PA, USA. Association for Computational Linguistics.

R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, and D. Tufis. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of LREC*.

J. Tiedemann and L. Lars Nygaard. 2004. The OPUS corpus - parallel & free. In *Proceedings of LREC*.

C. Tillmann and H. Ney. 2000. Word re-ordering and dp-based search in statistical machine translation. In *In Proc. of the COLING 2000, JulyAugust*, pages 850–856.