# Improving Speech Translation with Automatic Boundary Prediction

*Evgeny Matusov*[1], *Dustin Hillard*[2], *Mathew Magimai-Doss*[3],
*Dilek Hakkani-Tur*[3], *Mari Ostendorf*[2], *Hermann Ney*[1]

[1]Lehrstuhl fuer Informatik 6, RWTH Aachen University, Germany
[2]Electrical Engineering, University of Washington, Seattle, WA, USA
[3]International Computer Science Institute, Berkeley, CA, USA

## Abstract

This paper investigates the influence of automatic sentence boundary and sub-sentence punctuation prediction on machine translation (MT) of automatically recognized speech. We use prosodic and lexical cues to determine sentence boundaries, and successfully combine two complementary approaches to sentence boundary prediction. We also introduce a new feature for segmentation prediction that directly considers the assumptions of the phrase translation model. In addition, we show how automatically predicted commas can be used to constrain reordering in MT search. We evaluate the presented methods using a state-of-the-art phrase-based statistical MT system on two large vocabulary tasks. We find that careful optimization of the segmentation parameters directly for translation quality improves the translation results in comparison to independent optimization for segmentation quality of the predicted source language sentence boundaries.

## 1. Introduction

Machine translation of automatically recognized speech is currently an important research topic. Yet most state-of-the-art automatic speech recognition (ASR) systems were developed without considering the possible use of recognized word sequences as input to MT. They typically do not produce sentence-like units (SUs), nor predict punctuation marks. The recognized words are divided into *utterances* based on speech/non-speech detection algorithms. These utterances may be very long, containing several sentences, or very short sentence fragments (1-2 words). MT systems are often not able to translate (with an acceptable quality) utterances that are too long or too short.

In this work we investigate the influence of automatic sentence segmentation on MT quality. We compare and combine existing SU boundary detection algorithms [12, 3] and measure their performance by evaluating the translations which utilize these boundaries. The translations are produced by a state-of-the-art phrase-based statistical MT system. We also introduce a new feature for sentence segmentation that makes use of the phrase translation model from this MT system. Adding this feature gives better MT results despite lower F-scores for sentence prediction. Finally, we also present an approach for detection of sub-sentence boundaries, which mark possible clauses and may correspond to commas. We show that these "soft" boundaries can be used to constrain reordering in the MT search, while the phrasal context across these boundaries is nevertheless considered by the MT system.

This paper is organized as follows. In Section 2, we present the approaches to sentence segmentation. Our baseline MT system is presented in Section 3. In Section 4, we introduce an algorithm for predicting within-sentence boundaries and explain how these boundaries can be interpreted as constraints to reordering in MT search. Section 5 describes the boundary prediction experiments performed for the Chinese-to-English and Arabic-to-English large vocabulary translation tasks. We conclude with a summary in Section 6.

## 2. Sentence Segmentation

### 2.1. ICSI+ Approach

In this work, we use the ICSI+ multilingual sentence segmentation tools [12] for both comma and sentence boundary detection. The sentence boundary detection is treated as a binary classification problem, where every word boundary can be of one of two classes: sentence boundary or non-sentence boundary. The classifier uses a combination of 5-gram hidden-event language models (HELM) and a boosting classifier [8] that combines speaker, prosodic, and lexical cues. The prosodic features include various measures of pause duration, phone duration, fundamental frequency and energy, and their normalized versions. The posterior estimates from the outputs of the two classifiers are interpolated using weights optimized on a held-out data set.

For Arabic, in addition to the boosting classifier we also make use of a support vector machines (SVM) classifier. Similar to [1], the posteriors estimated from the combination of the posterior estimates of the two individual classifiers is then interpolated with the HELM posteriors. The SVM has exactly the same feature input as that of the boosting classifier.

### 2.2. RWTH Approach

In state-of-the-art approaches to SU boundary detection [9], the boundaries are determined by selecting only those positions for which the posterior probability of a sentence boundary exceeds a certain threshold. This means that although the segmentation granularity can be controlled, the length of a segment may take any value from 1 to several hundred words. This may pose a problem for machine translation. Many statistical machine translation algorithms are either inefficient or not applicable if the length of the input sentence (in words) exceeds a certain threshold $L$. Also, if a segment is too short (e.g. less than 3-4 words), important context information can be lost.

The RWTH sentence segmentation algorithm [3] was developed especially for the needs of machine translation. It also uses the concept of hidden events to represent the segment boundaries. A decision to place a segment boundary is made based on a log-linear combination of language model and prosodic features. However, in contrast to the ICSI+ approach, we restrict the minimum and maximum length of a segment (e. g. 4 and 60 words, respectively) and add an explicit segment length model. As a result, the decision criterion and the HMM-style search has to be modified to include explicit optimization over the previous segment boundary.

The main features used by the algorithm are a 4-gram hidden-event LM, a normalized pause duration feature, and an explicit sentence length probability distribution learned from the training data. A segment penalty is also used to additionally control the segmentation granularity. The scaling factors in the log-linear combination of these and other models are tuned on a development set.

Other features can be included in the log-linear model. In particular, for a hypothesized boundary, the posterior probability from the ICSI+ model can be used as an additional feature to improve the RWTH approach.

### 2.3. Phrase Coverage Feature

Another feature that can be included in the RWTH approach is motivated by the phrase-based machine translation algorithm that will be applied to the segmented speech in the next processing step. The idea is to make sure that word sequences for which good phrasal translations exist will not be broken into subsequences by a sentence boundary. To this end, we extract all bilingual phrases from the training data of the MT system (see Section 3) which match any word sequence in the evaluation data. Then, we train a bigram language model on the source language parts of these bilingual phrases. The phrases are treated as sentences, so words within the phrase (but not across phrases) are used to estimate the bigram.

The phrase coverage feature for each word $f_j$ in the input is then the bigram language model probability $p(f_{j+1}|f_j)$. If this probability is high, the word sequence $f_j f_{j+1}$ most probably has a good phrasal translation, and a sentence boundary directly after $f_j$ is undesirable. If this probability is low, the MT system will probably translate each of the two words by backing off to single-word translations. In this case, the phrasal context will be lost anyway, so that an (incorrect) boundary between $f_j$ and $f_{j+1}$ will not have a significant negative influence on translation quality. Note that by introducing the phrase coverage feature we may improve the MT quality, but not necessarily improve the segmentation results with respect to precision and recall.

## 3. Baseline MT System

In our experiments, we use a state-of-the-art phrase-based translation system [4]. In this system, a target language translation $e_1^I = e_1 \dots e_i \dots e_I$ for the source language sentence $f_1^J = f_1 \dots f_j \dots f_J$ is found by maximizing the posterior probability $Pr(e_1^I|f_1^J)$. This probability is modeled directly using a log-linear combination of several models. The best translation is found with the following decision rule:

$$\hat{e}_1^{\hat{I}} = \underset{I, e_1^I}{\operatorname{argmax}} \left\{ \sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J) \right\} \quad (1)$$

The model scaling factors $\lambda_m$ for the features $h_m$ are trained with respect to the final translation quality measured by an error criterion [5]. The baseline system includes an $n$-gram language model, a phrase translation model, and a word-based lexicon model. The latter two models are used for both directions: $p(f|e)$ and $p(e|f)$. Additionally, we use a word penalty and a phrase penalty. Other features include the phrase count thresholds and a phrase reordering model (see [4] and Section 4.2).

## 4. Soft Boundaries

Motivated by analysis of a small corpus of human word alignments where we found that very little reordering occurs across commas, we investigate the use of automatically predicted commas as soft boundary constraints for translation reordering.

### 4.1. Soft Boundary Prediction

In this work, commas are predicted using the same approach as in sentence boundary prediction (Section 2.1), employing the same lexical and prosodic features, with the exception of speaker change. While comma and sentence boundary prediction could be treated jointly as a multi-class problem, here we take predicted sentence boundaries as given and then predict commas (without distinguishing between comma and caesura) within the sentence.

### 4.2. Using Soft Boundaries in MT

One of the features in the loglinear translation model in Section 3 is the reordering model. The reordering model of the baseline system is a distance-based model. It assigns costs based on the distance from the end position of a phrase to the start position of the next phrase; "jumps" over a long distance are penalized. For Chinese-to-English translation, this simple model is combined with a maximum entropy model predicting the probability of a phrase orientation class [11].

In this work, we extend the reordering model by an additional penalty, the *soft boundary penalty*. Reordering across a soft boundary is assumed to be highly unlikely and is penalized. The soft boundaries described in Section 4.1 implicitly divide a source sentence into several parts. Each word $f_j$ at position $j$ in a sentence is labeled with an integer label $c(j)$ which encodes the (soft boundary separated) section of the sentence that the word is from. We penalize the movement of a phrase from the position $j$ to a position $j'$ by a weight $\alpha$ if the two positions have different section labels:

$$w(j, j') = \alpha \cdot |c(j') - c(j)| \quad (2)$$

The reason for introducing such a penalty is the assumption that the words between two soft boundaries usually represent a sentence clause. Nevertheless, the phrasal translation and language model context beyond the soft boundary can be taken into account. This context is lost if we translate each sentence part as if it were a separate sentence. Note that the penalty in Eq. 2 naturally increases in case the hypothesized phrase movement is across two, three, etc. boundaries, making reordering from the beginning to the end of a long sentence very unlikely.

Given a text or a speech transcript with sub-sentence punctuation, we can consider commas to be soft boundaries and define the labels $c(j)$ accordingly. In case of automatically predicted soft boundaries , we can use the posterior probability of a boundary to make the penalty dependent on the confidence with which the soft boundary was predicted. Incorporating soft boundary confidence scores is straightforward: the labels $c(j)$ in Eq. 2 are replaced by real values $r(j)$, which are computed recursively as follows:

$$r(j) = \begin{cases} 0, & \text{if } j = 0 \\ r(j-1), & \text{if } c(j) = c(j-1) \\ r(j-1) - \log p_{nb}(j) & \text{if } c(j) \neq c(j-1) \end{cases} \quad (3)$$

Here, $p_{nb}(j)$ is the posterior probability that the soft boundary *does not* appear between the words $f_{j-1}$ and $f_j$. If the new position $j'$ and the old position $j$ of the first word in a phrase are in the same sentence part, no penalty will be added, since $r(j) - r(j') = 0$.

## 5. Experimental Results

### 5.1. Corpus Statistics

The experiments were performed on the GALE Chinese-to-English and Arabic-to-English large vocabulary tasks. We evaluated the segmentation and translation quality on the automatically recognized broadcast news portion of the GALE MT 2006 evaluation data. The ASR output was generated by the SRI 2006 Mandarin and Arabic evaluation systems. The reference transcriptions of the Chinese evaluation data contain about 19K characters and 633 sentence units. The Arabic reference transcriptions contain about 12K words and 661 sentence units.

Table 1: Corpus statistics for the bilingual training data of the Chinese-to-English and Arabic-to-English MT systems (GALE large data track).

| | | Source | Target |
|---|---|---|---|
| Chinese to English | Sentence Pairs | 7M | |
| | Running Words | 199M | 213M |
| | Vocabulary Size | 223K | 351K |
| Arabic to English | Sentence Pairs | 4M | |
| | Running Words | 126M | 125M |
| | Vocabulary Size | 421K | 337K |

The MT systems were trained using the bilingual training corpora from LDC. The statistics of the training corpora are shown in Table 1. For tuning the boundary prediction parameters we used a held out part of TDT4 as a development set for Chinese and the BBN 2006 tune set for Arabic. The baseline RWTH MT systems were initially optimized on the NIST 2004 evaluation data and further adjusted to the speech input using the GALE 2006 tune sets for Arabic and Chinese.

The Mandarin ASR system has a character error rate (CER) of 5.6% for the extended 2006 development, and 17.8% for the MT 2006 Evaluation set. The Arabic system has a WER of 17.1% on the BBN 2006 Tune set, 19.4% on the BBN 2006 Development set, and 33.7% on the MT 2006 evaluation set.

### 5.2. Evaluation Criteria

The quality of comma and SU prediction was measured in terms of precision (P), recall (R), and F-measure in comparison with manual reference boundaries defined on correct transcriptions. For the evaluation, the predicted boundaries were inserted into the reference text based on the edit distance alignment.

The MT quality was determined using the well-established objective error measures BLEU [6] and TER [7]. We used the tool of [2] to determine the alignment with the multiple reference translations based on the word error rate and, using this alignment, to re-segment the translation output to match the number of reference segments. For the evaluation data, the error measures were calculated using 3 manually created reference translations.[1] For the speech development data, only single reference translations were available.

The MT evaluation was case-insensitive, with punctuation marks. The punctuation marks were predicted by the MT system that had been trained by removing punctuation marks from the source phrases, but leaving them in the corresponding target phrases. Thus, the decision to insert punctuation marks was made by using the translation model and the target language model. This type of punctuation prediction has been shown to have some advantages over predicting punctuation in the source language before the translation or in the target language after the translation [3].

### 5.3. Language Model Training

In all experiments, we use $n$-gram language models with modified Kneser-Ney smoothing as implemented in the SRILM toolkit [10]. The HELMs for sentence boundary prediction were trained with the same data sources as for training the Chinese ASR language models, including broadcast news speech transcripts, TDT2 and TDT3 text data, the Chinese Gigaword corpus, the Chinese portion of various news translation corpora, and web news data collections from NTU and CU. The boosting models are trained using the TDT4 corpus. The HELM for comma prediction is trained on the Chinese Gigaword corpus, where the training text has been stripped of all punctuation but comma, caesura, and sentence boundaries.

We used a 4-gram target language model for the Arabic-to-English translation and a 6-gram language model for Chinese-to-English translation in the search. They were trained on the English part of the bilingual training corpus and additional monolingual English data from the Gigaword corpus. The total amount of language model training data was about 600M running words.

[1] These are the original manual reference translations produced for the GALE evaluation by NVTC, on the basis of which the "gold standard" translation was created.

Table 2: Segmentation and translation results [%] for different sentence segmentation settings on the Chinese-to-English task.

| algorithm | P | R | F-score | BLEU | TER |
|---|---|---|---|---|---|
| ICSI+ 0.8 | 93.1 | 38.6 | 54.6 | 19.2 | 68.5 |
| ICSI+ 0.5 | 81.8 | 64.8 | 72.3 | 20.2 | 67.5 |
| ICSI+ 0.2 | 69.6 | 83.2 | 75.8 | 20.7 | 67.3 |
| RWTH | 72.2 | 74.3 | 73.2 | 20.7 | 67.4 |
| + phrase LM | 57.2 | 82.2 | 67.5 | 21.2 | 67.0 |
| RWTH+ICSI | 75.0 | 77.5 | 76.2 | 20.8 | 67.1 |
| boundary after every 30 words | | | | 18.1 | 69.7 |
| reference sentence units | | | | 20.7 | 66.9 |

### 5.4. MT Results for SU Boundary Prediction

Table 2 summarizes the segmentation and translation results for the ICSI+ and RWTH algorithms. In the ICSI+ approach, the boundaries are inserted if the sentence end posterior probability exceeds a certain threshold. Here, we tried the thresholds 0.2, 0.5, and 0.8, which led to average segment lengths of 16, 24, and 45 words, respectively. The best threshold determined on a development set is 0.2. This means that shorter segments are better for translation, i. e. recall is more important than precision. For this algorithm, the setting with the highest F-score also results in the best translation quality.

The RWTH approach has a lower F-score than ICSI+, but performs similarly in terms of BLEU and TER. One advantage of this algorithm is that extreme sentence lengths cannot occur in its output. Here, the minimum and maximum SU length was set to 4 and 60 words, respectively. In contrast, even using a low posterior probability threshold of 0.2 that favors short SUs, the ICSI+ system produced 5 sentences that were 100 or more words long. 40 sentences contained only 1 word. Most probably, the translations of these "sentences" were not adequate. Segmentation quality improves when ICSI posteriors are used as a feature in the RWTH approach (RWTH+ICSI), but translation quality is equivalent to the individual approaches.

The best translation quality (BLEU score of 21.2) is achieved by adding the phrase coverage feature described in Section 2.3. It is notable that the F-score for this setup is low, but the recall is high. The phrase coverage feature results in additional SU boundaries that may not correspond to manually defined boundaries, but have less impact on the translation because phrasal context at these extra boundaries was not captured during MT training.

For comparison, we also report the translation results for two baseline setups. In the first setup, a boundary is inserted after every 30 words in a document. This is clearly not a good idea, since the BLEU score is low. In the second setup, the manual reference boundaries are inserted into the ASR output based on the alignment with the correct transcriptions. We see that the automatic SU boundary prediction results in translations of the same or even somewhat better quality than when reference boundaries are used.

Table 3: Segmentation and translation results [%] for different sentence segmentation settings on the Arabic-to-English task.

| algorithm | P | R | F-score | BLEU | TER |
|---|---|---|---|---|---|
| ICSI+ 0.8 | 76.9 | 43.3 | 55.4 | 21.8 | 62.2 |
| ICSI+ 0.2 | 40.1 | 84.9 | 54.4 | 21.6 | 62.8 |
| RWTH | 52.6 | 54.4 | 53.5 | 22.0 | 62.3 |
| + phrase LM | 49.7 | 60.3 | 54.5 | 22.1 | 61.9 |
| RWTH+ICSI | 61.3 | 68.8 | 64.8 | 21.9 | 62.4 |
| boundary after every 30 words | | | | 20.6 | 63.7 |
| reference sentence units | | | | 21.5 | 62.4 |

Table 4: Comma and translation results [%] for the different SU and soft boundary settings on the Chinese-to-English task.

| SU algorithm | P | R | F-score | BLEU | TER |
|---|---|---|---|---|---|
| reference SUs | 100 | 100 | 100 | 20.8 | 66.9 |
| RWTH+ICSI | 73.8 | 35.6 | 48.0 | 20.7 | 67.1 |
| ICSI+ 0.5 | 77.0 | 40.1 | 52.7 | 20.3 | 67.4 |

In Table 3, we report the results for the same experiments on the Arabic-to-English task. Here, the F-measures for the SU boundaries are lower than for Chinese. The main difference relative to Chinese-to-English translations is that it is advantageous to produce longer segments (an ICSI threshold of .8 led to an average length of 33 words). We attribute this to the fact that reordering is mostly local when translating from Arabic to English. If two sentences are translated as one, their words are usually not swapped. In general, the Arabic-to-English MT is less sensitive to SU boundaries than the Chinese-to-English MT. All automatic segmentation approaches are as good in terms of MT quality as when the reference SU boundaries are inserted into the ASR output.

### 5.5. MT Results for Soft Boundary Prediction

The sentence segmentation results presented in Section 5.4 show that shorter segments can be better translated by the Chinese-to-English system than long segments. One reason for this is that erroneous reordering across a missed SU boundary can cause translation errors. However, the context information is often lost when short segments split sentences because each piece is then translated individually. So, especially for Chinese, the prediction of soft boundaries could constrain and thus correct MT reordering without the negative effect of cutting the context.

Table 4 presents the comma prediction and translation results for three settings. In the first setting, we used the integer penalties $c(j)$ as in Eq. 2. The penalties were computed relative to the reference (oracle) commas and SU boundaries that we inserted into the ASR output. The second setting uses automatically predicted commas and their posterior probabilities as in Eq. 3, which were inserted given the SU boundaries predicted by the RWTH+ICSI system. Here, we considered only the commas with probability $> 0.2$. In the third setting, we used the commas predicted given the somewhat longer SUs of the ICSI+ system at a threshold of $0.5$, which resulted in using more automatically predicted commas (with higher comma recall). Comma recall increases for less frequent sentence boundaries because inserted SUs can often occur at reference comma locations. In all cases, the BLEU and TER improvements were not significant with respect to the translation results in Table 2 without using the soft boundaries. We attribute this in part to the good quality of the baseline maximum entropy reordering model that already restricts unnecessary long-range phrase reordering. Nevertheless, in some translated sentences, word order and cause-effect relations were subjectively more correct when the soft boundary penalty was used. In the future, we plan to further analyze and improve the soft boundary concept.

## 6. Conclusion

In this work, we test the importance of segment boundaries in automatically recognized speech for MT quality. We combine two approaches for SU boundary prediction in order to produce sentences that are best suited for a state-of-the art phase-based statistical MT system. We used a novel feature, phrase coverage, in order to couple the segmentation with the predictive power of the phrase translation model. We also employed an automatic comma prediction algorithm and used the produced commas as soft boundaries, constraining reordering in the MT search. Our experiments find that the best translation results are achieved when boundary detection algorithms are directly optimized for translation quality.

## 7. Acknowledgments

Thanks to Richard Zens, Liz Shriberg, James Fung, Sebastien Cuendet, Yang Liu and Matthias Zimmerman for their contributions. This paper is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023.

## 8. References

[1] Magimai-Doss, M., et al. "Entropy based classifier combination for sentence segmentation," in *Proc. ICASSP*, 2007.

[2] Matusov, E., Leusch, G., Bender, O., and Ney, H. "Evaluating Machine Translation Output with Automatic Sentence Segmentation," in *Proc. of IWSLT 2005*, pp. 148-154, October 2005.

[3] Matusov, E., Mauser, A., and Ney, H. "Automatic Sentence Segmentation and Punctuation Prediction for Spoken Language Translation," in *Proc. of IWSLT 2006*, pp. 158–165, November 2006.

[4] Mauser, A., et al. "The RWTH Statistical Machine Translation System for the IWSLT 2006 Evaluation," in *Proc. of IWSLT 2006*, pp. 103–110, November 2006.

[5] Och, F. J. "Minimum error rate training in statistical machine translation," in *Proc. of the ACL 2003*, pp. 160–167, July 2006.

[6] Papineni, K., et al. "BLEU: A method for automatic evaluation of machine translation," in *Proc. of the ACL 2002*, pp. 311–318, July 2002.

[7] Snover, M., Dorr, B., Schwartz, R., Micciula, L., and Makhoul, J. "A study of translation edit rate with targeted human annotation," in *Proc. of AMTA 2006*.

[8] Schapire, R. E., Singer, Y. "Boostexter: A boosting-based system for text categorization," *Machine Learning*, vol. 39, no. 2/3, pp. 135–168, 2000.

[9] Shriberg, E., Stolcke, A., Hakkani-Tur, D., and Tur, G. "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Communication*, vol. 32, no. 1-2, pp. 127–154, 2000.

[10] Stolcke, A. "SRILM – an extensible language modeling toolkit," in *Proc. ICSLP 2002*, volume 2, pp. 901–904.

[11] Zens, R., and Ney, H. "Discriminative Reordering Models for Statistical Machine Translation," in *HLT-NAACL: Proc. of the Workshop on Statistical Machine Translation*, pp. 55–63, June 2006.

[12] Zimmerman, M., et al. "The ICSI+ multilingual sentence segmentation system," in *Proc. Interspeech*, 2006.