CROSS-LINGUAL SPEECH-BASED TOBI LABEL GENERATION USING BIDIRECTIONAL LSTM

Marco Vetter* Sakriani Sakti† Satoshi Nakamura†

* Nara Institute of Science and Technology, Japan
†RIKEN, Center for Advanced Intelligence Project AIP, Japan

{marco.vetter.mp8, ssakti, s-nakamura}@is.naist.jp

ABSTRACT

In this paper we investigate the automatic generation of ToBI-style prosody labels. The work is motivated by the idea of using prosodic information to facilitate the automatic lexicon discovery for unseen and under-resourced languages for which sufficient training data is not available. Specifically, the prosodic boundaries are meant to serve as additional top-down information in the word segmentation step. To this end we attempt to apply the trained Japanese models cross-lingually on a language not seen in training (English). We generate break index labels, using only the speech signal as input, with no additional information given at test time in the form of transcripts or prior word segmentations. The labels are generated using bidirectional LSTMs trained on spontaneous Japanese speech. We evaluate the quality of these labels using established metrics, with an F1 score of 0.55 for cross-lingual prosodic break detection (given a tolerance of 80 ms).

Index Terms—Prosody detection, ToBI label generation, cross-lingual speech processing, word segmentation

1. INTRODUCTION

There are currently over 7000 living languages in the world [1], many of which are only spoken by small and often shrinking groups of speakers and are therefore threatened by extinction [2]. A central task in documenting these languages is word discovery, which is preceded by the step of segmenting the speech signal into word-candidate segments.

Language documentation is a time-consuming task, making the process ultimately expensive. Natural Language Processing (NLP) systems could be useful tools to facilitate the exploration and documentation of previously unseen languages. Unfortunately, such systems are often reliant on large quantities of annotated data, which is generally not available for smaller languages. One strategy to circumvent this issue is to exploit the similarities between languages by training a system on one or more well-resourced languages and then applying it to the under-resourced target language.

Infants have been shown to recognize prosody before acquiring the ability to segment speech into smaller segments such as words and clauses [3] [4]. We therefore will attempt to generate prosodic boundaries, which could then be used as additional information for the word segmentation process.

Specifically, our goal is to create prosodic break index labels as introduced in [5] (for English) and [6] [7] (for Japanese). We intend to achieve this by applying neural network models in a cross-lingual fashion in order to create these labels for a language not seen in training. Notably, we will only use the speech signal as input when applying the model at test time; no additional information is provided to the system in form of transcriptions or existing segmentations of any kind.

2. RELATED WORK

Both, the effects of applying prosodic information to the problem of word segmentation and automatically generating ToBI style prosody labels have been explored in past research. In [8], [9] Ludusan et al. have shown that using oracle prosodic information as well as boundaries detected based on acoustic cues can improve the performance of term discovery.

As for generating ToBI and other prosodic labels, in [10] Syrdal et al. found ToBI labels predicted from text were able to speed up labelling by humans, another potential application of automatically generated ToBI labels. In [11] Chen et al. were able to generate a simplified set of prosodic labels using ANN- and GMM-based models that use phoneme transcriptions and acoustic observations, with prosody-dependent pronunciations pre-compiled in a lexicon. Rosenberg has published his AuToBI system for automatic ToBI annotation in [12]. AuToBI uses speech recordings and TextGrid files with existing word segmentations as input to produce ToBI labels for English speech. In [13], [14] the same system was used for cross-language prominence and boundary detection. Similarly, the Eti.ToBI tool published in [15] by Elvira-Garcia uses a speech wave form and a TextGrid file containing syllable boundaries and marks for lexically stressed syllables. With these it generates labels according to the Sp.ToBI and
Japanese speech to train a neural network. The similarities in the way prosody is expressed across languages, the system will be able to use the knowledge gained from one language to segment speech from another language with similar degrees of success.

As for the model, we choose bidirectional long short-term memory neural networks (BiLSTM). These have been shown to work well on time-series labelling tasks, such as boundary detection [17], [18]. The network is relatively light-weight, consisting of two hidden layers with 1024 BiLSTM cells each.

### 4. DATA

#### 4.1. Corpora

The Corpus of Spontaneous Japanese (CSJ) [19] contains a large amount of annotated recordings, a sub-set of which also offers ToBI-style labels as defined in [7]. For our experiments we will use the “academic” and “simulated public speech” parts of the prosody-annotated CSJ, which amounts to approximately 38 hours of data.

The Boston University Radio News Corpus (BRC) is split into two parts, radio and lab news. The lab news part of the corpus consists of a sub-set of the original radio broadcast stories, re-recorded in a laboratory, amounting to a total of approx. 78 minutes of speech.

#### 4.2. ToBI Prosodic Break Index Labels

The main set of basic labels introduced in the J_ToBI standard consists of four break index levels ranging from 0 (strong cohesion) to 3 (strong degree of disjuncture) [6]. X-JToBI extends this set with various modifiers for boundaries that lie between these levels. It also adds additional types of labels and modifiers for word fragments, word-internal pauses and prosodic filler [7]. Since many label types and modifiers are very rare in the data, training robust models on them is not feasible. We will therefore rewrite the labels before training in two different ways.

The first set of experiments will be conducted using a simple binary mapping, after which we will run a second set of experiments using all of the basic labels (without modifiers). For the multi-class experiments, we are mapping the English ToBI labels to their respective counterparts according to table 1.

### 5. EXPERIMENTS

#### 5.1. Network architecture and features

For our experiments we are using a relatively light-weight BiLSTM consisting of two hidden layers with 1024 BiLSTM cells
cells each, implemented with the PyTorch deep learning platform. Activations and loss calculation use the tanh and cross-entropy loss functions, respectively.

Due to the continuous nature of the speech and the relative similarity of features in neighbouring frames, predicting frame-exact boundaries is extremely difficult. For this reason, when attempting to evaluate boundary detection, the system is often granted a tolerance. For phoneme segmentation tolerances of 20 to 40 ms have been used in the past [20][21][22]. Since the time scale for prosodic events is larger than that for phonemic and sub-phonemic segmentation, we will generally report scores for tolerances of 40 and 80 ms, but also for exact matches (0 ms).

Another problem when training neural networks for this kind of task is the extreme skewness of the data. The vast majority of frames does not represent a boundary (98.6%). To counteract this, PyTorch allows declaring weights for individual classes, which are applied during loss calculation.

Finally, the system tends to produce clusters of break labels around those time indices it believes to be prosodic boundaries. This is likely due to the similarity of feature vectors representing neighbouring frames of the speech signal. Until we devise a way to prevent this behaviour, we will apply post-processing to the network output by reducing any label clusters to the central time index of that cluster.

As for feature extraction, we used the KALDI speech recognition toolkit [23] to extract MFCCs with a 25 ms window and 10 ms frame shift. After adding deltas and delta-deltas the process resulted in 39-dimensional feature vectors.

### 5.2. Results on Japanese data

First we applied the model trained on binary labels (boundary / no boundary) to Japanese data also taken from the CSJ. Results are shown in table 2.

As we can see, tolerance significantly impacts the scores. Exact matches (0 ms tolerance) are extremely rare, as shown by the low precision, and the system only catches approx. 20% of all the boundaries in the reference. But even a tolerance similar to that used in phoneme recognition (40 ms) yields much better results, with 44.9% of predicted boundaries within four frames of a true boundary. At 80 ms we are able to find 61.37% of all boundaries in the ground truth.

We can also see that the biggest improvements in scores take place until around 30-40 ms of tolerance. Giving the system more leeway than that still results in additional predicted boundaries being classified as correct, but the vast majority are within 30-40 ms of a reference boundary.

Next we trained a network to perform a multi-class labelling task. The results for 80 ms of tolerance can be found in table 3. Obviously there are vast differences in performance with regard to the various break label types. Word fragments (D) and word internal pauses (P) have proven very difficult. It should be noted that these are also the two least frequent label types in the data. Of the disfluencies, the prosodic Filler (F) was the easiest to detect. Level 2 breaks (accentual phrase) were the most difficult of the main types. Level 1 prosodic breaks (AP-medial word boundaries) show the highest scores, followed by level 3 breaks (intonation phrase).

### 5.3. Results on English data

For comparison, we also trained monolingual English systems on the majority of the BRC lab news data (~65 minutes), referred to as "BRC" in tables below. We then applied these systems and the Japanese ones to English test data. Results for binary labels are shown in table 4.

<table>
<thead>
<tr>
<th>Break type</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.4947</td>
<td>0.6272</td>
<td>0.5504</td>
</tr>
<tr>
<td>2</td>
<td>0.3306</td>
<td>0.1620</td>
<td>0.2114</td>
</tr>
<tr>
<td>3</td>
<td>0.4723</td>
<td>0.3770</td>
<td>0.3991</td>
</tr>
<tr>
<td>D</td>
<td>0.5601</td>
<td>0.0113</td>
<td>0.0161</td>
</tr>
<tr>
<td>F</td>
<td>0.3993</td>
<td>0.2000</td>
<td>0.2514</td>
</tr>
<tr>
<td>P</td>
<td>1.0000</td>
<td>0.0705</td>
<td>0.0705</td>
</tr>
</tbody>
</table>

Table 3. Results for multi-class labels on CSJ data (for 80 ms tolerance)

<table>
<thead>
<tr>
<th>System</th>
<th>Tolerance (ms)</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRC</td>
<td>0</td>
<td>0.0751</td>
<td>0.0931</td>
<td>0.0825</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>0.4730</td>
<td>0.5934</td>
<td>0.5226</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>0.6308</td>
<td>0.7880</td>
<td>0.6956</td>
</tr>
<tr>
<td>CSJ</td>
<td>0</td>
<td>0.0716</td>
<td>0.0636</td>
<td>0.0673</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>0.3963</td>
<td>0.3510</td>
<td>0.3716</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>0.5914</td>
<td>0.5216</td>
<td>0.5533</td>
</tr>
</tbody>
</table>

Table 4. Results for binary labels on BRC data

The scores show that on this task the cross-lingual system loses between ~20 and 30% performance (as indicated by F1 score), depending on tolerance. But it is able to detect more than half of the prosodic boundaries in the English test data.
within a tolerance of 80 ms, without ever having been exposed to English speech in training. Compared to the monolingual results presented in table 2, we see that frame-exact performance is noticeably worse. However, as we increase tolerance, scores improve drastically, so that even in cross-lingual application 59.14% of all predicted boundaries fall within 80 ms of a true boundary.

Finally, we applied the multi-class model to the BRC data, results for which (at 80 ms tolerance) can be found in table 5.

<table>
<thead>
<tr>
<th>System</th>
<th>Break type</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRC</td>
<td>1</td>
<td>0.5238</td>
<td>0.3614</td>
<td>0.4177</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.6330</td>
<td>0.0137</td>
<td>0.0225</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.7031</td>
<td>0.1700</td>
<td>0.2555</td>
</tr>
<tr>
<td>CSJ</td>
<td>1</td>
<td>0.3816</td>
<td>0.6578</td>
<td>0.4765</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.1279</td>
<td>0.1299</td>
<td>0.1229</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.2204</td>
<td>0.1974</td>
<td>0.2050</td>
</tr>
</tbody>
</table>

Table 5. Multi-class results on BRC data (80 ms tolerance)

The X-JToBI labels for disfluencies do not exist in the English ToBI annotations and are therefore not part of these results. As in the binary case, compared to the monolingual Japanese results from table 3, scores are overall lower. This is to be expected considering the increased difficulty inherent in cross-lingual model application. But especially the scores for type 1 breaks are close to the performance reported on Japanese. Also, the cross-lingual system actually performs better for some break types (1, 2) when compared to the monolingual English system.

5.4. Discussion

We have trained BiLSTMs on Japanese speech with ToBI-style prosodic break annotations. Due to the relative rarity of some labels we merged infrequent types, resulting in one binary and one multi-class mapping.

We then applied the trained models to both Japanese and English data, and compared cross-lingual results with a monolingual English system. The binary model was able to detect prosodic breaks on Japanese data with some accuracy. Most of the correctly predicted boundaries fell within 30-40 ms of their respective reference label. The same model applied to English data performed noticeably worse with regard to frame-exact matches. However, given a reasonable tolerance, performance was close to that on Japanese data, and within ~20-30% of the monolingual English system.

The multi-class boundary detector performed with varying degrees of success on the different break label types. Especially labels very rarely seen in the data proved difficult to detect. As for cross-lingual multi-class detection, the system trained on a large amount of Japanese data performed better with regard to more subtle break types (1, 2), while the monolingual system trained on little English data performed better detecting type 3 breaks. Although this may be caused by the limited amount of training data available in the BRC (~65 minutes) compared to the CSJ (~37 hours), it nevertheless shows that the fundamental approach of cross-lingual prosodic boundary detection is valid, and can be applied to situations where sufficient data to train monolingual models is not available.

Figure 1 shows a visualization of reference and cross-lingually generated ToBI labels for English speech using the speech analysis tool Praat.

![Fig. 1. ToBI break labels for English. (a) speech signal, (b) spectrogram, (c) reference labels, (d) cross-lingual labels](image)

We can see that the first four type 1 reference labels are correctly identified, although the system places them slightly earlier than the human annotator did. The final type 3 label is also correctly detected, with the system placing it a short time after the reference. It is also apparent that the system tends to produce labels where the reference does not feature them at all, leading to the reduced precision scores we have seen earlier.

6. CONCLUSION AND FUTURE WORK

In this paper we have attempted to use neural networks in cross-lingual application in order to predict prosodic boundaries on a language not seen in training. We have shown that the chosen model does retain much of its predictive power in cross-lingual application. If we can improve the overall system, we would expect an increase in monolingual performance to carry over to cross-lingual application.

Feature extraction may be a point at which improvements could be possible, e.g. by using more prosody-specific features. We may also be able to expand our approach to ToBI-style intonation labels, making for a complete automatic ToBI-labelling system for Japanese and potentially cross-lingual application. The main goal remains to use the generated prosodic information to improve the results of word segmentation algorithms to aid automatic lexical discovery.

7. ACKNOWLEDGEMENTS

Part of this work was supported by JSPS KAKENHI Grant Numbers JP17H06101 and JP17K00237.
8. REFERENCES


