USING PROSODY TO SPOT LOCATION MENTIONS

Gerardo Cervantes, Nigel G. Ward
University of Texas at El Paso

ABSTRACT
Identifying location mentions in speech is important for many information retrieval and information extraction tasks. Most commonly this is done with speech recognition and a gazetteer, but here we explore the value of prosody for this task. While previous work has explored the use of prosody for spotting named entities, including locations, the specific value of prosody for finding locations in spontaneous speech is not known. Using the Switchboard corpus and LSTM modeling we obtain above baseline performance. Further, we identify specific prosodic features and configurations that tend to mark locations in American English.

Index Terms— named entities, information retrieval, spontaneous speech, prosodic patterns, LSTM

1. INTRODUCTION
Location spotting is important for many tasks, including information retrieval, information extraction, question answering, summarization and translation. For speech, location spotting with speech recognition and a gazetteer is not always applicable and effective. First, for many low-resource languages there are no good recognizers, or no recognizers at all [1]. Second, even when good speech recognizers exist, many locations will be out-of-vocabulary, making the recognizer unable to find the location. Even without speech recognition, it can be useful to identify likely location mentions, either to send them to a human for transcription and lookup, or for special processing. For example, since location names tend to be pronounced similarly across languages, — for example Graz in English and Guraatsu in Japanese — cross-language ASR using acoustic models trained on other languages [2], and gazetteers in other languages may be effective.

For these reasons we are interested in ways to find locations without use of speech recognition. We hypothesize that prosody can be useful for this. Casual observation suggests that, across languages, introductory mentions of new entities, including locations, may share common prosodic features, such as late pitch peak. To the extent that locations are mentioned in certain specific contexts and associated with certain specific pragmatic functions, for example, introducing new topics or grounding, it makes sense that certain specific prosodic patterns may co-occur. Thus we thought that it may be possible to identify such general patterns, and then leverage information across languages.

2. RELATED WORK
In general, words of different classes and with different functions may have different typical prosodic forms. This has been exploited, for example, in spotting important words to include in summaries [3]. This general property of lexical prosody has also been exploited in language models which incorporate prosodic information [4, 5, 6]. More specifically, the value of prosodic information for named entity recognition has been shown, especially for person names.

Hakkani-Tur and colleagues did the first study of this [7], motivated by the idea that name mentions would generally have “prominent” prosody. For broadcast news, comparing with an entity tagger that used lexical information alone, they reported only a modest performance benefit, and found that the benefit came largely from distinguishing content words from function words, rather than from distinguishing entity mentions from other content words.

Work by Katerenchuk and Rosenberg [8], on the Wall Street Journal corpus, showed that acoustic (prosodic) cues can help detect named entities, when used in combination with recognizer output, in cases where the recognition error rate was high.

Rangarajan and Narayanan [9] obtained good results using prosody for detecting person names, although their task was relatively easy because they used read inputs, word boundaries were given, all input sentences contained exactly one person mention, and all person names were from a non-English language, but embedded in an English sentence.

Thus previous work has not shown whether prosody is useful for discriminating location mentions rather than just named entities in general, or even whether prosody is serving only to enable a general discrimination between content and function words. Previous work has also been limited to read speech; here we also examine the prosody of locations in spontaneous speech.

3. TASK
Our task is to identify likely places in speech where locations are being said. Classical formulations of the task of named
entity recognition assume that transcripts are available and exact word boundaries are important [10], which is not realistic in general. Instead, we formulate the task as one of identifying speech frames that have location mentions. Specifically, we aim to classify each point in the audio as part of a location mention (1) or not (0). In real-world applications, such labels would probably be smoothed or otherwise post-processed, however this task formulation is adequate for our aim here, namely to evaluate the pure ability of prosody to discriminate location mentions from all other speech regions.

4. DATA

We used the Switchboard corpus of American English telephone conversations, as this is large and fully transcribed with exact word boundaries [11, 12]. The locations are however not labeled in the transcripts. To find locations from the transcripts, we used spaCy, a natural language processing library. SpaCy has multiple downloadable neural network models that identify named entity types. When the spaCy classified a word with the entity types of GPE or location, these were marked as locations in the transcripts. The locations labeled by the spaCy model were not exact. For example, the word *Dallas in the Dallas Cowboys* was tagged as a location mention, although this word here is not a location but part of the team name. (However, depending on the intended purpose [13], spotting the word *Dallas* in this context as a location could still be useful.)

To evaluate the quality of these location tags, we hand-labeled the first 100 location mentions in 16 Switchboard conversations. Of these, 86 were tagged as locations by spaCy; thus the recall was 86.0%. In a sample of 98 words tagged as locations by spaCy, there were 12 false positives, and thus the precision was 87.7%. Thus these labels were only slightly noisy, so we chose to use them uncorrected, both for purposes of training and evaluation.

5. PROSODIC FEATURES

Our hypothesis is that, in addition to lexical and phonetic indications of locations, the prosody is also informative. For the two models described below, different feature sets were used. For the linear regression model, we accordingly use a wide set of prosodic and associated features, including not only track-normalized pitch, intensity, and duration, but also energy flux, and measures of the degree of creaky voice, lengthening, disalignment between intensity and pitch peaks, and the voiced/unvoiced intensity ratio. These were designed to be robust, as is necessary for spontaneous speech in general, and especially for Switchboard, given its varied audio quality [14]. Like other feature sets [15], this feature set has been shown in previous work to be informative regarding many semantic and pragmatic functions [14, 16, 17, 18].

<table>
<thead>
<tr>
<th></th>
<th>Linear Regression</th>
<th>LSTM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Threshold</td>
<td>0.329</td>
<td>0.033</td>
</tr>
<tr>
<td>Precision</td>
<td>0.532</td>
<td>0.532</td>
</tr>
<tr>
<td>Recall</td>
<td>0.95</td>
<td>0.945</td>
</tr>
<tr>
<td>F1-measure</td>
<td>0.682</td>
<td>0.681</td>
</tr>
</tbody>
</table>

Table 1. Model comparison on balanced datasets

Thinking that indications of location mention may be found not only on the word itself or its immediate neighbors, we used prosodic features spanning a wide context, extending 3200 milliseconds before and after the frame to be classified. Thinking that the behavior of the interlocutor may also be informative, we used prosodic features for both speakers. We computed features over fixed-length windows, without concern for alignment to word, utterance, or syllable boundaries, as we cannot in general assume that these are available.

Further, for the LSTM models, described below, we used a reduced feature set, since LSTMs are in general able to learn temporal patterns, such as the dynamics of and relations among pitch and intensity. LSTMs have shown to require few frame-level prosodic features to achieve good results [19]. For the LSTM, we accordingly used only 5 features per speaker, each computed frame-by-frame, namely absolute pitch, z-normalized pitch, voicing, energy, and cepstral flux (as an indicator for both rate and phonetic reduction). Each frame in the audio thus had 10 (5 + 5) prosodic features.

The code for computing these prosodic features is available open-source in the Mid-Level Prosodic Feature Toolkit [20].

6. MODELS AND TRAINING

We experimented with two models: linear regression, because it is easy to analyze what it learns, and a Long Short Term Memory (LSTM) model, because it can learn temporal patterns and has demonstrated good performance in numerous speech processing tasks.

For both models, 15% of the data was used for testing, 15% for dev, and the rest was used as training data. Since the predictions given by the models are continuous-valued, they were converted to binary by using a threshold. The threshold was set to the value that gave the highest performance on the dev dataset by the F1-measure. This threshold was then used for the test set for evaluation.

6.1. Linear Regression Model

Location mentions are not that common: only 1 in 256 frames have locations in this data. To enable learning in linear regression, we accordingly downsampled to have equal numbers of positive and negative examples. Specifically, all frames that had a location mention are used, and the negative frames were
Table 1 compares the performance of the linear regression and LSTM models. Both were evaluated on evenly balanced data, and non-speech frames were excluded. However the data was not exactly the same, the non-speech frames were different as they were randomly selected with a different random seed. For the linear regression model, we downsampled the negative frames, as described above. The LSTM Model had to be tested on 10-second segments, and it made a prediction for each frame in each segment, but before computing precision and recall we downsampled the negative-class frames so that the data was balanced in this case also.

We see that both models have higher precision than baseline (0.50) and that the linear regression performs just slightly better than the LSTM model.

7.2. Comparison to Baselines

To understand the level of performance for the LSTM, Table 2 shows results for three baselines: a) Random baseline: we wanted see if the model was doing better than random. b) Speaking baseline: we were interested in finding if the model was doing better than a baseline that has perfect knowledge of whether there is speech or not. This baseline predicts randomly only when there is speech, as given by the transcripts. c) Content-Word baseline: we were interested in a smarter baseline that has knowledge of whether there is speech, as well as knowledge of whether the word being said was a function word or a content word. Function words are used to express grammatical relationships and can not have locations. We defined function words to be those on the NLTK stoplist. Thus the baseline only predicted randomly when there were content words, and predicted false otherwise, as per the transcripts.

Further, to evaluate whether the interlocutor-track features were informative, we built another LSTM model using only one track, excluding features computed from the audio track of the other speaker. As seen in the rightmost column of Table 2, the performance was slightly lower, indicating that interlocutor features might have a low impact on performance.

7.3. Locations and Other Entities

We next sought to determine whether our model had indeed learned to spot location frames, rather than frames with entities. We hypothesized that the prediction values for frames that were locations would be generally higher than the prediction values for other named entities. We wrote a script to gather all capitalized words; these were in general names of people and organizations, and we used this set, uncorrected,
Table 3. Precision of trained English conversation model evaluated over different datasets

<table>
<thead>
<tr>
<th></th>
<th>English News</th>
<th>English Conversation</th>
<th>Spanish Conversation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech Baseline</td>
<td>2.5%</td>
<td>0.8%</td>
<td>0.9%</td>
</tr>
<tr>
<td>Content Baseline</td>
<td>3.0%</td>
<td>–</td>
<td>1.0%</td>
</tr>
<tr>
<td>Model</td>
<td>9.0%</td>
<td>3.0%</td>
<td>3.0%</td>
</tr>
</tbody>
</table>

7.4. Generality Across Languages and Genres

As a preliminary investigation of whether the model was specific to this language and this data set, we did some small-scale experimentation with other data sets. Since we did not have timestamped transcripts for any of these, our evaluation was done in a post hoc fashion, based on examination of timepoints for which our model had high location estimates. We started from the highest likelihood frame and worked down the list. However, as high-estimate frames tended to be clustered in time, to get a more diverse sampling, we skipped over frames within one second of those already examined. For each language, we examined the top 100 timepoints the LSTM model predicted in this way and computed the precision.

For comparison, we annotated randomly selected points in the audio until we found 100 random timepoints that had non-function words. For the speech-only baseline, laughter, music and silence timepoints were excluded. In each case, the precision was computed by taking the number of locations found over the number of timepoints examined. To enable comparison, we also examined 100 predictions for Switchboard in the same way. The first comparison dataset was an English news broadcast dataset: 6 hours of local news broadcasts data from different stations [18]. As these had only a single audio track, we used the single track model as seen in Table 2. As seen in columns 1 and 2 of Table 3, there appear to be many more locations in this data, and the model appears useful for identifying them. The other dataset was the Spanish Callhome telephone conversation corpora, approximately 10 hours of data. As seen in Table 3, the model results were slightly above baseline. This suggests that prosody of locations could have similarities across languages.

8. FEATURE ANALYSIS

To get a rough idea of how prosody was enabling detection of locations, we inspected the coefficients of correlation with the presence or absence of a location frame (1/0). The first finding was that the correlations were time-dependent. For example, intensity correlated positively with upcoming frames being locations, but negatively with recent past frames being locations. Table 4 shows all features whose correlation’s absolute value was greater than 0.02, ordered by time: the times are the window starts and ends relative to the frame being classified. All correlations shown were significant (p < 10^{-12}).

<table>
<thead>
<tr>
<th>feature</th>
<th>window (ms)</th>
<th>correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>speaker pitch narrowness</td>
<td>–800 ~ –400</td>
<td>+0.24</td>
</tr>
<tr>
<td>speaker pitch narrowness</td>
<td>–400 ~ –200</td>
<td>+0.20</td>
</tr>
<tr>
<td>speaker intensity</td>
<td>–200 ~ –000</td>
<td>–0.26</td>
</tr>
<tr>
<td>speaker lengthening</td>
<td>–300 ~ –200</td>
<td>+0.26</td>
</tr>
<tr>
<td>speaker pitch narrowness</td>
<td>0 ~ +200</td>
<td>–0.25</td>
</tr>
<tr>
<td>speaker intensity</td>
<td>+600 ~ +800</td>
<td>–0.24</td>
</tr>
</tbody>
</table>

Table 4. Prosodic features best correlating with location mentions.

9. CONCLUSIONS AND FUTURE WORK

We have shown that prosodic information is useful for spotting location mentions. Further, this ability is somewhat location-specific, beyond any generic benefit of being able to distinguish content words from function words, and even beyond any generic ability to spot entity mentions.

Based on very small samples, the performance of an English-trained model appears low on other languages, but within English appears to generalize to the news genre.

Future work might explore the possible value of partly shared network training and the presence of possible universals. Future work should also quantify the extent to which the information provided by prosody is a useful (non-redundant) complement to that provided by speech recognizers, for languages for which those are available.

10. ACKNOWLEDGEMENTS

We thank: Olac Fuentes for discussion, Isabel Ward and Aaron Alarcon for helping with the post hoc evaluation. This work was supported in part by DARPA’s LORELEI program, but no official endorsement should be inferred.
11. REFERENCES


