Vol.2018-SLP-125 No.17
2018/12/10

Automatic Prediction of Symbolic and Sentence-Level Prosody in English for Development of a Reading Tutor

Xinyi Zhao[1,a)] Nobuaki Minematsu[1,b)] Daisuke Saito[1,c)]

Abstract: In English education, speech synthesis technologies can be effectively used to develop a reading tutor to show students how to read given sentences in a natural and native way. The tutor can not only provide native-like audio of the input sentences but also visualize required prosodic structure to read those sentences aloud naturally. As the first step to develop such a reading tutor, prosodic events that can imply the intonation of the sentence need to be predicted from plain text. In this research, phrase boundary and 4-level stress instead of the traditional binary stress level are taken into consideration as prosodic events. 4-level stress labels not only categorize syllables into stressed ones and unstressed ones, but also indicate where phrase stress and sentence stress should appear in a sentence. Conditional Random Fields as a popular sequence labeling method are employed to do the prediction work. Experiments showed that applying our proposed method can improve the performance of prosody prediction compared to previous researches.

Keywords: English prosody, CRFs, phrase boundary, stress level

1. Introduction

Prosody is an important component employed to express emotions and intentions in speaking. It can influence the speech understanding because it usually contains information beyond literal meaning of a sentence. Most of the students may find it difficult to study English prosody due to the lack of effective methods, and unnatural prosody makes learners difficult to be understood, and of course, sound much less native. Therefore, to help with this situation, automatic prediction of prosodic events from input English text is needed when developing a reading tutor.

There are mainly two aspects of prosody we are concerned with in this paper, the phrase boundary and stress. The prosodic phrase boundary, which is often represented by "slash," divides a sentence into phrasal segments, and can usually decide where should pause in that sentence. The stress, on the other hand, refers to which syllable in a word should be pronounced to be salient.

Since the automatic prediction of prosodic events started to be pursued technically, several machine learning and deep learning methods have been proven to be effective in such a task, including Support Vector Machine (SVM) [1], Conditional Random Fields (CRFs) [2, 3], Hidden Markov Model (HMM) [4], and BLSTM-RNN [5,6]. When stress prediction was involved, most researches only used binary annotation for stress, where syllables were categorized into stressed ones and unstressed ones. However, binary stress level is not adequate enough to guide English learners.

In this paper, we proposed to employ CRFs to predict the phrase boundary and stress from plain text to develop a reading tutor. We mainly focused on the prediction of stress on the phrase level and stress on the sentence level, which were seldom distinguished from each other in prior researches.

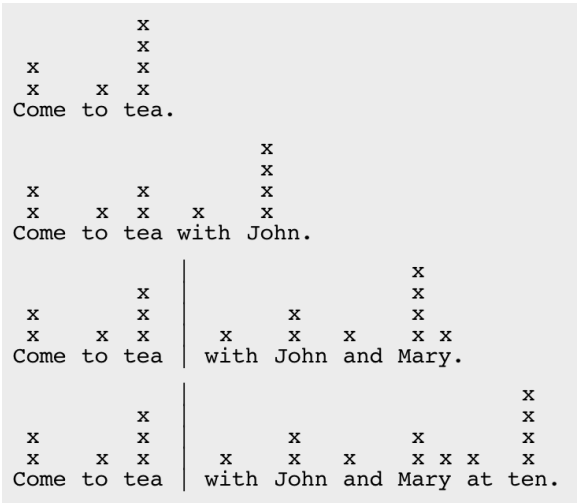

Fig. 1 Stress in a sentence

2. Phrase Stress and Sentence Stress

Fig. 1 shows an example that can explain the stress inside a sentence [7]. The number of "x" marks on each syllable from one to four stands for unstressed syllable, word stress, phrase stress and sentence stress, respectively.

In the first sentence "Come to tea," the prominent word is "tea," which is also called sentence stress in our work. While in sentence "Come to tea with John," the sentence stress falls on the word "John," showing that when the sentence changes, the sentence stress can be transferred onto another word. After this sen-

[1] The University of Tokyo, Bunkyō, Tokyo 113-8656, Japan
[a)] xinyi@gavo.t.u-tokyo.ac.jp
[b)] mine@gavo.t.u-tokyo.ac.jp
[c)] dsk_saito@gavo.t.u-tokyo.ac.jp

tence being extended into "Come to tea with John and Mary," the native speaker will naturally divide it into two phrases. In the first phrase "Come to tea," like the first simple sentence, the prominent word is "tea," but compared to the prominent word in the second phrase, it no longer receives the biggest emphasis in the whole sentence. We call the prominent word in a phrase as phrase stress. Obviously, there is only one word per phrase that may have phrase stress, and only one word per sentence that may have sentence stress, and when a word has sentence stress, it must also has the phrase stress.

Just as mentioned in a classic linguistic research by Liberman [8], "the strongest stress in a phrase will fall as far back, i.e., as close to the end as possible," which is also called nuclear stress rule that linguists use to annotate the stress on words and syllables theoretically. To assign the "x" marks like Fig. 1, some more detailed rules were summarized [7]. Each syllable is assigned with one "x" at first, then at word level, every polysyllabic word and monosyllabic content word gets marked with an additional "x." At the phrase level, only one word in a phrase gets marked with an extra "x," which is usually the last content word inside the phrase. Similarly, at the sentence level, a single word gets marked with an extra "x," which is usually the last content word in the sentence. When applying these rules, however, there are many exceptions. For example, in the phrase "for the first time," the stress tends to fall on the word "first" instead of "time," and in the phrase "orange juice," the first noun receives more emphasis than the last noun. These exceptions show that simple rules are unable to give the right answer all the time, and more efforts should be paid.

3. Tasks and Corpus

When English learners are not sure how to pronounce a certain word, they can always look up a dictionary and it will offer information of phoneme and stress position inside this word. But things are different when they try to pronounce a complete sentence. Though pronouncing every word correctly, English learners may still sound strange, and what makes it worse is that there is no dictionary to tell them where to pause or where to place stress in a sentence. To help with this situation, we aim at developing a reading tutor that can automatically predict the position of phrase boundaries, phrase stress and sentence stress.

In this study, we employed the machine learning method, CRFs, to do the prediction work. CRFs, as a popular and efficient method for a sequence labeling task, have been successfully used in many automatic prosodic annotation works [2, 3, 9]. Though the applied corpora are various and it is unfair to judge methods only by the results of these researches, the results can indeed reflect how well these methods performed in prosodic annotation tasks to some extent. CRFs gave one of the best results in phrase boundary prediction with recall rate being 0.679, precision being 0.753, and f-score being 0.714 [2], which were very close to results of SVM. While in the binary prediction of whether a word is a prominent word, the results given by CRFs exceeded others significantly, with recall rate being 0.950, precision being 0.927 and f-score being 0.939 [3]. Having had successful experience in prior researches, it is reasonable to employ CRFs in our work.

Table 1 Basic features for CRFs

1	Token
2	Base form of token
3	POS

The data we used to train and test are provided by an expert phonetician who has rich experience in teaching English to Japanese learners. The corpus consists of 712 carefully chosen English sentences, each of which is recorded in North American accent. The corresponding text was analyzed prosodically and manually, and was annotated with phrase boundaries and the stress level on each syllable. The stress level system was formed by 4 levels, including level one representing unstressed syllable, level two representing stressed syllable inside a word, level three standing for phrase stress and level four being sentence stress. Fig. 1 are some examples of her prosodic annotations.

4. Experiments

Both experiments of symbolic prediction of phrase boundaries and stress level were conducted. In phrase boundary prediction, the performance of CART in Festival Speech Synthesis System was evaluated as baseline, for it is one of a common and popular tool in this task. As for stress level prediction, only the experiments applying CRFs were conducted, because we did not find any system or prior research that realized 4-level stress prediction. All the recall and precision of CRFs are the average of 10-fold cross validation, with the proportion of training data and testing data being 9 to 1.

4.1 Phrase boundary

The phrase boundary prediction can be formulated as a problem of after which word in a given sentence a slash representing the boundary should be inserted. For each word in the sentence, the judgment should be done and the binary result of inserting slash or not can be given.

As mentioned above, we firstly conducted the predicting experiment as baseline by using Festival which is the default tool for predicting phrase boundaries in the HMM/DNN-based Speech Synthesis System (HTS). Festival offers example modules for speech synthesis, for example, the part-of-speech (POS) tagging for each word and duration prediction for each phoneme. For phrase boundary prediction, there are two methods provided, classification and regression trees (CART) and probabilistic model, and here as baseline method, we chose the simpler one, CART. Results are shown in Table 3 for comparison, and just as the document of Festival suggests, the baseline method was trained to predict the boundary with rare false detection, causing low recall but high precision.

The position of phrase boundaries has strong correlation with the syntactic and lexical structure of a sentence, which can be excellent references when choosing features for the CRFs. Refering to a prior research [2], we listed a very basic set of features that should be extracted from previous, current and next tokens, as shown in Table 1.

Then an improved feature set inspired by some prior researches [2, 3, 9] that also employed CRFs to predict phrase boundaries.

One of them originally predicted prosodic events in Japanese text which does not even have obvious boundary between words, and we found it could be a good reference for a similar task in English. As shown in Table 2, in this improved feature set, more efforts were put on the context by adding bigram and trigram features, with letter "C" standing for current, "P" for previous and "N" for next.

The results of both experiments applying the basic features and the improved features are listed in Table 3. It shows that CRFs indeed performed better than the baseline method, and with appropriate features chosen, the result can be improved further.

By analyzing the predicted results which were not matched with the source data, we found that quite a few of them were actually acceptable speaking style for a reading tutor, which means the position of phrase boundaries in source data are not the only correct position to pause in a sentence. For example, in the sentence "Peter apologized for his temper and his impatience," our prediction claims that there is a phrase boundary behind the word "apologized," while the source data claims there is not, but both are judged as acceptable by a phonetician.

4.2 Stress

Experiments of predicting on which word phrase stress or sentence stress falls were conducted. To the best of our knowledge, unlike predicting phrase boundaries, an existing tool for phrase and sentence stress prediction was not found, and as a result, there is no baseline method in this task.

With the experience of slash position prediction, we have summarized a list of features that was effective in predicting phrase boundaries, as shown in Table 2. Considering that the stress position can also be closely correlated with these features, we firstly applied the same features for phrase and sentence stress prediction, and the results are respectively shown in Table 5 and Table 6 for comparison.

Besides the above features, another important feature that can effect the phrase stress position is actually the phrase boundary, the one we just predicted in the other experiment. As mentioned in section 2, linguistic rules for stress assignment show how important the position of phrase boundary is to phrase stress, and to verify whether this theory can also influence the prediction using CRFs, an experiment of predicting phrase stress only with the features listed in Table 4 was conducted.

The results listed in Table 5 show the experiment applying correct phrase boundary as a feature performed much better than applying the long list of features of Table 2, suggesting that phrase boundary is one of the most important features for phrase stress prediction.

However, the phrase boundary is not given from plain text data, and what we have is just the predicted phrase boundary. Assuming that our prediction result is accurate enough, we conducted another experiment using the predicted phrase boundary instead of the correct one, with two other features, the token and POS, staying the same. The results are also listed in Table 5. It shows that applying predicted phrase boundary performed better than using the improved set of features, but the recall, precision and f-score are much lower than applying the correct phrase bound-

Table 2 Improved features for CRFs

Features of P, C, N tokens individually	
1	Token
2	Base form of token
3	POS
4	Lexical stress pattern (according to CMU pronouncing dictionary)
N-Gram features	
5	Bigram features of P and C tokens
6	Bigram features of C and N tokens
7	Trigram features of P, C and N tokens

Table 3 Results of phrase boundary prediction

Method	Recall	Precision	F-score
CART	0.290	0.787	0.424
CRFs (basic features)	0.686	0.748	0.715
CRFs (improved features)	0.712	0.761	0.735

Table 4 Features for phrase and sentence stress prediction

1	Token
2	POS
3	Correct slash position
4	Word's backward position in the phrase

Table 5 Results of phrase stress prediction

Method	Recall	Precision	F-score
CRFs (improved features)	0.802	0.806	0.804
CRFs (correct slash)	0.915	0.922	0.919
CRFs (predicted slash)	0.801	0.822	0.811

Table 6 Results of sentence stress prediction

Method	Recall	Precision	F-score
CRFs (improved features)	0.857	0.875	0.865
CRFs (correct slash)	0.870	0.892	0.881
CRFs (predicted slash)	0.862	0.890	0.876

ary, suggesting that improvement is needed in the prediction of phrase boundary in the future.

Similarly, the experiments of sentence stress prediction were also conducted, both using the correct slash labels and predicted ones. In addition, the phrase's backward position in the whole sentence was applied as a new feature. The results are listed in Table 6. Comparing to the long list of features, when phrase boundary is involved, it is impressive that fewer features can actually provide better performance, showing that phrase boundary is also important for sentence stress prediction.

The same fact that the annotation in source data is not the only correct answer to the prosody of the sentence is also true in stress prediction task. For example, in the sentence "only the most accomplished artists obtain popularity," whether the phrase boundary falls on the word "accomplished" as in source data or on "artists" as in prediction, they are both acceptable. Reevaluation of the obtained results not using source data but using judgements of other phoneticians will be done in the future.

5. Conclusions

In this work, we applied CRFs to predict the phrase boundary, phrase stress and sentence stress from plain text. It was shown that using CRFs can indeed provide good performance in predicting prosodic events. Choosing the appropriate features for CRFs can improve the performance further. In phrase stress and sentence stress prediction, it was proven that the position of phrase stress and sentence stress is closely correlated with the position of phrase boundary, so the phrase boundary should be chosen as an important feature.

The prediction results are to some extent satisfactory, but improvement is also needed. We can either improve our feature set for CRFs prediction, or employ neural-network (NN) based method on the same tasks. However, to apply NN based method, we need a larger corpus than the one we used in this research. Together with the fact that the corpus we used was not designed for educational purposes, we are planning to enlarge the size of training data by annotating an additional text corpus. One possibility is use of sentences in extensive reading. We have seleted several candidates, but considering that manual annotation can be very time-consuming, another possibility is to use the CMU_ARCTIC database. For each sentence in this CMU database, the tone tier in ToBI labeling and some phrase boundary related labels are provided. If we can find a proper way to convert these labels to the format we proposed, the size of training data can then be easily enlarged.

References

[1] 行野顕正他, "統計的アプローチによる英語スラッシュ・リーディング教材の自動生成", 情報処理学会論文誌, 48(1), 365–374, 2007.

[2] 永田亮, 樽谷久翔, "品詞解析/統語解析を必要としない英語スラッシュ・リーディング教材自動生成手法", 電子情報通信学会論文誌 D, 95(2), 264–274, 2012.

[3] Nagata, R. *et al.*, "A Method for Predicting Stressed Words in Teaching Materials for English Jazz Chants," *IEICE Transactions on Information and Systems*, 95(11), 2658–2663, 2012.

[4] Nagata, R. *et al.*, "Toward a chanting robot for interactively teaching English to children," *Second Language Studies: Acquisition, Learning, Education and Technology*, 2010.

[5] Zhao, Y. *et al.*, "A study on BLSTM-RNN-based Chinese prosodic structure prediction in a unified framework with character-level features," *Proc. Speech Prosody*, 64–68, 2016.

[6] Rendel, A. *et al.*, "Using continuous lexical embeddings to improve symbolic-prosody prediction in a text-to-speech front-end," *Proc. ICASSP*, 5655–5659, 2016.

[7] Erickson, D, "Jaw movement and rhythm in English dialogues," *Technical Report of Acoust. Soc. Jpn.*, H-98-59, 1–8, 1998.

[8] Liberman, M., Prince, A, "On stress and linguistic rhythm," *Linguistic inquiry*, 8(2), 249–336, 1977.

[9] 鈴木雅之他, "条件付き確率場を用いた日本語東京方言のアクセント結合自動推定", 電子情報通信学会論文誌 D, 96(3), 644–654, 2013.