Predicting Prosodic Prominence from Text with Pre-trained Contextualized Word Representations

Aarne Talman,[*] [†] **Antti Suni,**[*] **Hande Celikkanat,**[*] **Sofoklis Kakouros,**[*]
Jörg Tiedemann[*] and **Martti Vainio**[*]
[*]Department of Digital Humanities, University of Helsinki, Finland
[†]Basement AI, Finland
{name.surname}@helsinki.fi

Abstract

In this paper we introduce a new natural language processing dataset and benchmark for predicting prosodic prominence from written text. To our knowledge this will be the largest publicly available dataset with prosodic labels. We describe the dataset construction and the resulting benchmark dataset in detail and train a number of different models ranging from feature-based classifiers to neural network systems for the prediction of discretized prosodic prominence. We show that pre-trained contextualized word representations from BERT outperform the other models even with less than 10% of the training data. Finally we discuss the dataset in light of the results and point to future research and plans for further improving both the dataset and methods of predicting prosodic prominence from text. The dataset and the code for the models are publicly available.

1 Introduction

Prosodic prominence, i.e., the amount of emphasis that a speaker gives to a word, has been widely studied in phonetics and speech processing. However, the research on text-based natural language processing (NLP) methods for predicting prosodic prominence is somewhat limited. Even in the text-to-speech synthesis domain, with many recent methodological advances, work on symbolic prosody prediction has lagged behind. We believe that this is mainly due to the lack of suitable datasets. Existing, publicly available annotated speech corpora, are very small by current standards.

In this paper we introduce a new NLP dataset and benchmark for predicting prosodic prominence from text which is based on the recently published LibriTTS corpus (Zen et al., 2019), containing automatically generated prosodic prominence labels for over 260 hours or 2.8 million words of English audio books, read by 1230 different speakers. To our knowledge this will be the largest publicly available dataset with prosodic annotations. We first give some background about prosodic prominence and related research in Section 2. We then describe the dataset construction and annotation method in Section 3.

Prosody prediction can be turned into a sequence labeling task by giving each word in a text a discrete prominence value based on the amount of emphasis the speaker gives to the word when reading the text. In Section 4 we explain the experiments and the experimental results using a number of different sequence labeling approaches and show that pre-trained contextualized word representations from BERT (Devlin et al., 2019) outperform our other baselines even with less than 10% of the training data. Although BERT has been previously applied in various sequence labeling tasks, like named entity recognition (Devlin et al., 2019), to the best of our knowledge, this is the first application of BERT in the task of predicting prosodic prominence. We analyse the results in Section 5, comparing BERT to a bidirectional long short-term memory (BiLSTM) model and looking at the types of errors made by these selected models. We find that BERT outperforms the BiLSTM model across all the labels.

Finally in Section 6 we discuss the methods in light of the experimental results and highlight areas that are known to negatively impact the results. We also discuss the relevance of pre-training for the task of predicting prosodic prominence. We conclude by pointing to future research both in developing better methods for predicting prosodic prominence but also to further improve the quality of the dataset. The dataset and the PyTorch code for the models are

Proceedings of the 22nd Nordic Conference on Computational Linguistics (NoDaLiDa), pages 281–290
Turku, Finland, 30 September – 2 October,

available on GitHub: `https://github.com/Helsinki-NLP/prosody`.

2 Background

2.1 Prosodic Prominence

Every word and utterance in speech encompasses phonetic and phonological properties that are not resulting from the choice of the underlying lexical items and that encode meaning in addition to that of the individual lexemes. These properties are referred to as prosody and they depend on a variety of factors such as the semantic and syntactic relations between these items, and their rhythmic grouping (Wagner and Watson, 2010). Prosodic variation in speech contributes to a large extend to the perception of natural sounding speech. Prosodic prominence represents one type of prosodic phenomenon that manifests through the subjective impression of emphasis in speech where certain words are interpreted as more salient within their lexical surrounding context (Wagner and Watson, 2010; Terken and Hermes, 2000).

Due to the inherent difficulty in determining prominence — even for human subjects, see, e.g., (Yoon et al., 2004) — the development of automatic tools for the annotation of prominent units has been a difficult task. This is exemplified from the large degree of discrepancy observed between human annotators when labeling prominence where the inter-transcriber agreement can vary substantially based on a multitude of factors such as the choice of annotators or annotation method (Mo et al., 2008; Yoon et al., 2004; Kakouros and Räsänen, 2016). Similarly, in prominence production, certain degree of freedom in prominence placement and large variability between styles and speakers (Yuan et al., 2005), renders the task of prominence prediction from text very difficult compared to most NLP tasks involving text only.

2.2 Generating Prominence Annotations

Throughout the literature a number of methods have been proposed for the labeling of prosodic prominence. These methods can be roughly categorized on the basis of the need for training data (manual prosodic annotations) into supervised and unsupervised, but crucially, on the basis of the information they utilize from speech and language to generate their predictions (prominence labels). As prominence perception has been found to correlate with acoustic-phonetic features (Lieberman, 1960), with the constituent syntactic structure of an utterance (Gregory and Altun, 2004; Wagner and Watson, 2010; Bresnan, 1973), with the frequency of occurrence of individual lexical items (Nenkova et al., 2007; Jurafsky et al., 2001), and with the probabilities of contiguous lexical sequences (Jurafsky, 1996), automatic methods have been developed utilizing these features either in combination or independently (Nenkova et al., 2007; Kakouros et al., 2016; Ostendorf et al., 1995; Levow, 2008).

Overall, these features can be largely divided into two categories: (i) *acoustic* (derived from the sound pressure waveform of the speech signal) and (ii) *language* (extracted by studying the form of the language; for instance, semantic or syntactic factors in the language). Both acoustic and language-based features have been shown to provide good overall performance in detecting prominence (in both supervised and unsupervised cases), where, however, the methods utilizing acoustic features seem to provide better performance for the unsupervised detection of prominences in speech (Suni et al., 2017; Wang and Narayanan, 2007; Kakouros and Räsänen, 2016), with state-of-the-art results reaching high level of accuracy, close to that of the inter-annotator agreement for the data. While the top-down linguistic information is known to correlate with perceptual prominence, in this paper we want to make a clear distinction between data labelling and text-based prediction. Thus, in this work, we utilize purely acoustic prominence annotations of the speech data using the method developed by Suni et al. (2017) as the prosodic reference.

2.3 Predicting Prosodic Prominence from Text

To what extent prosodic prominence can be predicted from textual input only has been a topic of inquiry in linguistics for a long time. In traditional generative phonology (Chomsky and Halle, 1968), accent placement was considered to be fully determined by linguistic structure, whereas a seminal work by Bolinger (1972) emphasized the importance and relevance of the lexical semantic context as well as the speakers' intention, positing that, in general, a mind reading ability may be necessary to determine prominent words in a sentence.

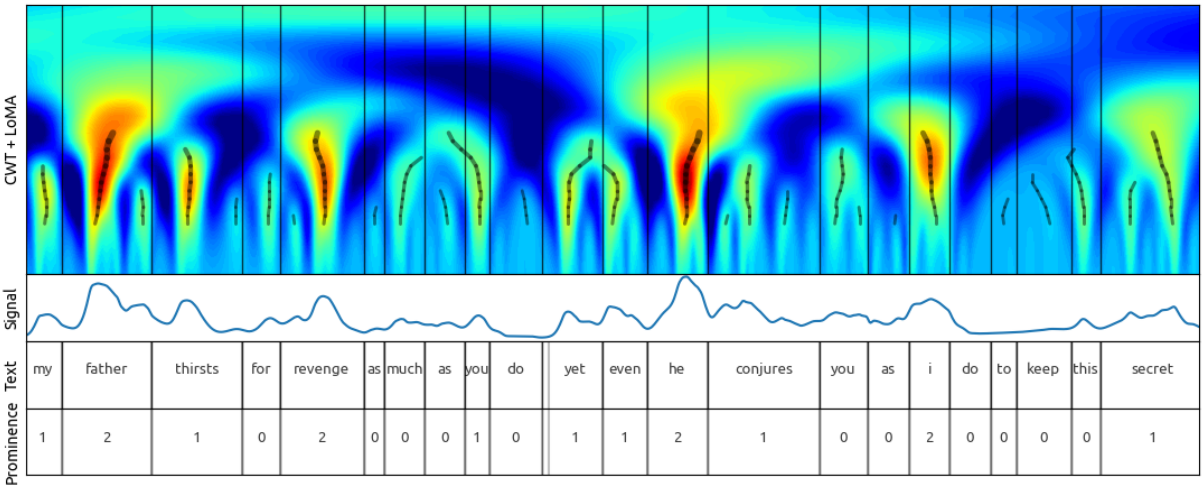

Figure 1: Continuous Wavelet Transform Annotation method.

				non-prominent	prominent	
sets (clean)	**speakers**	**sentences**	**words**	**0**	**1**	**2**
train-100	247	33,041	570,592	274,184	155,849	140,559
train-360	904	116,262	2,076,289	1,003,454	569,769	503,066
dev	40	5,726	99,200	47,535	27,454	24,211
test	39	4,821	90,063	43,234	24,543	22,286
total:	1230	159,850	2,836,144	1,368,407	777,615	690,122

Table 1: Dataset statistics

As longstanding inquiries hold, the goal of reliably predicting the placement of prominent entities from information automatically derived from textual resources is still ongoing.

Several efforts have been made towards this direction, especially in text-to-speech (TTS) synthesis research, where generation of appropriate prosody would increase both intelligibility and quality of synthetic speech. Before the deep learning paradigm shift in NLP, several linguistic features were examined for prominence prediction, including function-content word distinction, part-of-speech class, and information status (Hirschberg, 1993). Statistical features like unigrams, bigrams, and TF-IDF have also been frequently used (Marsi et al., 2003). Later, the accent ratio, or simply the average accent status of a word type in the given corpus, was found to be a stronger predictor than linguistic features in the accent prediction task (Nenkova et al., 2007), suggesting that lexical information may be more relevant than linguistic structure for the prominence prediction task.

Recently, continuous representations of words have become commonplace in prosody prediction for TTS, though the symbolic level is often omitted and pitch and duration are predicted directly using lexical embeddings (Watts, 2012). Yet, closely related to the proposed method, (Rendel et al., 2016) experimented with various lexical embeddings as an input to a Bi-directional LSTM model, predicting binary prominence labels. Training on a proprietary, manually annotated single speaker corpus of 3730 sentences, they achieved an F-score of 0.71 with Word2Vec (Mikolov et al., 2013) embeddings, with a clear improvement over traditional linguistic features.

3 Dataset

We introduce, automatically generated, high quality prosodic annotations for the recently published LibriTTS corpus (Zen et al., 2019). The LibriTTS corpus is a cleaned subset of LibriSpeech corpus (Panayotov et al., 2015), derived from English audiobooks of the LibriVox project.[1] We selected the 'clean' subsets of LibriTTS for annotation, comprising of 262.5 hours of read speech from 1230 speakers. The transcribed sentences were aligned

[1] https://librivox.org

Token	Tell	me	you	rascal	,	where	is	the	pig	?
Discrete label	2	0	0	0	NA	2	0	0	1	NA
Real-valued label	1.473	0.333	0.003	0.167	NA	2.160	0.006	0.037	0.719	NA

Table 2: Example sentence with the annotation from the dataset. Discrete prominence values were used in the experiments of this paper. The real-valued labels are used for generation of the discrete labels, however, they could also be used directly for prominence prediction.

with the Montreal forced aligner (McAuliffe et al., 2017), using a pronunciation lexicon and acoustic models trained on the LibriSpeech dataset. The aligned sentences were then prosodically annotated with word-level acoustic prominence labels. For the annotation, we used the Wavelet Prosody Analyzer toolkit[2], which implements the method described in (Suni et al., 2017). Briefly, the method consists of 1) the extraction of pitch and energy signals from the speech data and duration from the word level alignments, 2) filling the unvoiced gaps in extracted signals by interpolation followed by smoothing and normalizing, 3) combining the normalized signals by summing or multiplication, and 4) performing a continuous wavelet transform (CWT) on the composite signal and extracting continuous prominence values as lines of maximum amplitude across wavelet scales (see Figure 1). Essentially, the method assumes that the louder, the longer, and the higher, the more prominent. On top of this, the wavelet transform provides multi-resolution contextual information; the more the word stands out from its environment in various time scales, the more prominent the word is perceived.

For the current study, continuous prominence values were discretized to two (non-prominent, prominent) or three (non prominent, somewhat prominent, very prominent) classes. The binary case is closely related to the pitch accent detection task, aiming for results comparable with the majority of the literature on the topic. The weights in constructing the composite signal and discretization thresholds were adjusted based on The Boston University radio news corpus (Ostendorf et al., 1995), containing manually annotated pitch accent labels. This corpus is often used in the evaluation of pitch accent annotation and prediction quality, with the current annotation method yielding state-of-the-art accuracy in word level acoustic-based accent detection, 85.3%, using weights 1.0, 0.5 and 1.0 for F0, energy and duration respectively, and using multiplication of these features in signal composition. For three-way discretization, the non-prominent / prominent cut-off was maintained and the prominent class was split to two classes of roughly equal size. Statistics of the resulting dataset are described in table 1. The full dataset is available for download here: `https://github.com/Helsinki-NLP/prosody`. Although not discussed in this paper, the described acoustic annotation and text-based prediction methods can be applied to prosodic boundaries too, and the boundary labels will be included in the dataset at a later stage.

4 Experiments

In this section we describe the experimental setup and the results from our experiments in predicting discrete prosodic prominence labels from text using the corpus described above.

4.1 Experimental Setup

We performed experiments with the following models:

- BERT-base uncased (Devlin et al., 2019)
- 3-layer 600D Bidirectional Long Short-Term Memory (BiLSTM) (Hochreiter and Schmidhuber, 1997)
- Minitagger (SVM) (Stratos and Collins, 2015) + GloVe (Pennington et al., 2014)
- MarMoT (CRF) (Mueller et al., 2013)
- Majority class per word

The models were selected so that they cover a wide variety of different architectures from feature-based statistical approaches to neural networks and pre-trained language models. The models are described in more detail below.

We use the Huggingface PyTorch implementation of BERT available in the `pytorch_transformers` library,[3] which

[2] `https://github.com/asuni/wavelet_prosody_toolkit`

[3] `https://github.com/huggingface/`

we further fine-tune during training. We take the last hidden layer of BERT and train a single fully-connected classifier layer on top of it, mapping the representation of each word to the labels. For our experiments we use the smaller BERT-base model using the uncased alternative. We use a batch size of 32 and fine-tune the model for 2 epochs.

For BiLSTM we use pre-trained 300D GloVe 840B word embeddings (Pennington et al., 2014). The initial word embeddings are fine-tuned during training. As with BERT, we add one fully-connected classifier layer on top of the BiLSTM, mapping the representation of each word to the labels. We use a dropout of 0.2 between the layers of the BiLSTM. We use a batch size of 64 and train the model for 5 epochs.

For the SVM we use Minitagger[4] implementation by Stratos and Collins (2015) using each dimension of the pre-trained 300D GloVe 840B word embeddings as features, with context-size 1, i.e. including the previous and the next word in the context.

For the conditional random field (CRF) model we use MarMot[5] by Mueller et al. (2013) with the default configuration. The model applies standard feature templates that are used for part-of-speech tagging such as surrounding words as well as suffix and prefix features. We did not optimize the feature model nor any of the other hyper-parameters.

All systems except the Minitagger and CRF are our implementations using PyTorch and are made available on GitHub: `https://github.com/Helsinki-NLP/prosody`.

For the experiments we used the larger train-360 training set. We report both 2-way and 3-way classification results. In the 2-way classification task we take the three prominence labels and merge labels 1 and 2 into a single prominent class.

4.2 Results

All models reach over 80% in the 2-way classification task while 3-way classification accuracy stays below 70% for all of them. The BERT-based model gets the highest accuracy of 83.2% and 68.6% in the 2-way and 3-way classification tasks, respectively, demonstrating the value of a pre-trained language model in this task. The 3-layer BiLSTM achieves 82.1% in the 2-way classification and 66.4% in the 3-way classification task.

The traditional feature-based classifiers perform slightly below the neural network models, with the CRF obtaining 81.8% and 66.4% for the two classification tasks, respectively. The Minitagger SVM model's test accuracies are slightly lower than the CRF's with 80.8% and 65.4% test accuracies. Finally taking a simple majority class per word gives 80.2% for the 2-way classification task and 62.4% for the 3-way classification task. The results are listed in Table 3. The fairly low results across the board highlight the difficulty of the task of predicting prosodic prominence from text.

To better understand how much training data is needed in the two classification tasks, we trained selected models with different size subsets of the train-360 training data. The selected subsets were: 1%, 5%, 10%, 50% and 100% of the training examples (token-label pairs). Figures 2 and 3 contain the learning curves for the 2-way and 3-way classification tasks, for all the models except for the majority and random baselines.

For all models and for both of the classification tasks we notice that they achieve quite high test accuracy already with a very small number of training examples. For most of the models the biggest improvement in performance is achieved when moving from 1% of the training examples to 5%. All models have reached close to their full predictive capacity with only 10% of the training examples. For example, BERT achieves 2-way classification test accuracy of 82.6% with 10% of the training data, which is only -0.6% points lower than the accuracy with the full training set. In the 3-way classification task 10% of the training data gives 67.1% for BERT, which is -1.7% points below the accuracy with the full training set.

Interestingly, in the 2-way classification task the BiLSTM model shows a slightly different learning curve, having already quite a high performance with just 1% of the training data, but then making no improvement between 1% and 5%. However, between 5% and 100% the BiLSTM model improvement is almost linear.

As the proposed dataset has been automatically generated as described in Section 3, we also tested the best two models, BERT and BiLSTM, with a manually annotated test set from The Boston University radio news corpus (Ostendorf et al., 1995).

`pytorch-transformers`

[4] `https://github.com/karlstratos/minitagger`

[5] `http://cistern.cis.lmu.de/marmot/`

Model	Test accuracy (2-way)	Test accuracy (3-way)
BERT-base	**83.2%**	**68.6%**
3-layer BiLSTM	82.1%	66.4%
CRF	81.8%	66.4%
SVM+GloVe	80.8%	65.4%
Majority class per word	80.2%	62.4%
Majority class	52.0%	48.0%
Random	49.0%	39.5%

Table 3: Experimental results (%) for the 2 and 3-way classification tasks.

For this experiment we trained the models using the train-360 training set (as above) replacing only the test set. The results of this experiment are shown in Table 4. The good results[6] from this experiment provide further support for the quality of the new dataset. Notice also that the difference between BERT and BiLSTM is much bigger with this test set (+3.9% compared to +1.1%). This difference could be due to the genre difference between the two test sets, with the Boston University news corpus being more contemporary compared to the source for our proposed dataset (pre-1923 books). This point will be further discussed in Section 6.

Model	vs expert	vs acoustic
BERT-base	**82.9%**	**82.1%**
3-layer BiLSTM	79.0%	79.3%

Table 4: Test accuracies (%) for the Boston University radio news corpus (2-way classification). expert = expert annotated perceptual prominence labels, acoustic = our acoustic prominence labels

5 Analysis

The experimental results show that although predicting prosodic prominence is a fairly difficult task, pre-trained contextualized word representations clearly help, as can be seen from the results for BERT. The difference between BERT and the other models is clear if we compare the other models with BERT fine-tuned with a small fraction of the training data. In fact, BERT already outperforms the other models with just 5% of the training examples in the 2-way classification case and with 10% of the training data in the 3-way classification

[6]Better results have been reported on Boston dataset using lexical features, but there are methodological concerns related to cross-validation training and speakers reading the same text, see discussion on (Rosenberg, 2009).

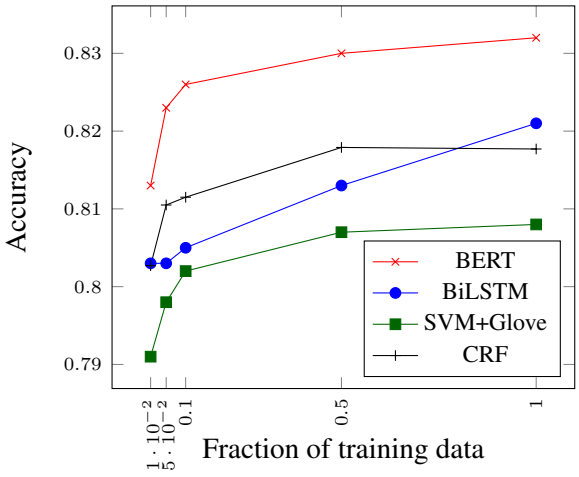

Figure 2: Test accuracy with different size subsets of the training data for the 2-way classification task.

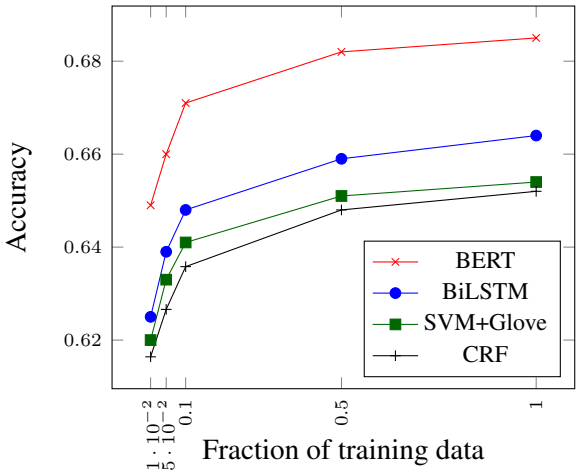

Figure 3: Test accuracy with different size subsets of the training data for the 3-way classification task.

case. This can be seen as an indication that BERT has acquired implicit semantic or syntactic information during pre-training that is useful in the task of predicting prosodic prominence.

To gain a better understanding of the types of predictive errors BERT makes, we look at the confusion matrices for the two classification tasks and compare those with the confusion matrices for the BiLSTM.

The 3-way classification confusion matrices are more informative as they allow comparison of the two models with respect to the predicted label in cases of error. Figure 4 contains the 3-way classification confusion matrix for BERT and Figure 5 for the BiLSTM model.

		Predicted			
		0	1	2	*recall*
Gold	0	35567	5602	2043	**82.3%**
	1	5943	11589	6987	**47.3%**
	2	1661	6208	14374	**64.6%**
	precision	**82.4%**	**49.5%**	**61.4%**	

Figure 4: 3-way classification task confusion matrix for BERT.

		Predicted			
		0	1	2	*recall*
Gold	0	35321	6157	1734	**81.0%**
	1	6221	12275	6019	**46.4%**
	2	2058	8014	12172	**61.1%**
	precision	**81.7%**	**50.1%**	**54.7%**	

Figure 5: 3-way classification task confusion matrix for BiLSTM.

In the 3-way classification task, when the gold label is 0 (non prominent) BERT makes more errors with prediction being 2 (very prominent) compared to the BiLSTM model. However, when the gold label is 2 (very prominent) BiLSTM makes more predictions with 0 (non prominent) compared to BERT. In general for 0 labels BERT seems to have higher precision and BiLSTM better recall, whereas for label 2 BERT has clearly higher recall and precision. Both models have low precision and recall for the less distinctive prominence (label 1). It seems that the clearest difference between the two models is in their ability to predict high prominence (label 2).

We also provide the confusion matrices for the 2-way classification task for the two models. Figure 6 contains the 2-way classification confusion matrix for BERT and Figure 7 for the BiLSTM model. Here BERT has slightly higher precision and recall across both of the labels.

		Predicted		
		0	1	*recall*
Gold	0	34249	8963	**79.3%**
	1	6186	40579	**86.8%**
	precision	**84.7%**	**81.9%**	

Figure 6: 2-way classification task confusion matrix for BERT.

		Predicted		
		0	1	*recall*
Gold	0	33786	9428	**78.2%**
	1	6670	40090	**85.7%**
	precision	**83.5%**	**81.0%**	

Figure 7: 2-way classification task confusion matrix for BiLSTM.

6 Discussion

We have shown above that prosodic prominence can reasonably well be predicted from text using different sequence-labelling approaches and models. However, the reported performance is still quite low, even for state-of-the-art systems based on large pre-trained language models such as BERT. We list a number of reasons for these shortcomings below and discuss their impact and potential mitigation.

Although the annotation method has been shown to be quite robust, errors in automatic alignment, signal processing, and quantization introduce noise to the labels. This noise might not be detrimental to the training due to dataset size, but the test results are affected. To measure the size of this effect, manual correction of a part of the test set could be beneficial.

It is well known that different speakers have different accents, varying reading proficiency, and reading tempo, which all impact the consistency of the labeling as the source speech data contains in total samples from over 1200 different speakers.

REF:	**One** way led to the left and the other to the right **straight** up the mountain .
BERT:	**One** way led to the **left** and the **other** to the right **straight** up the mountain .
REF:	In the **next** moment he was **concealed** by the leaves .
BERT:	In the **next** moment he was **concealed** by the leaves .
REF:	I **had** to read it over **carefully** , as the **text** must be absolutely correct .
BERT:	I **had** to read it over **carefully** , as the **text** must be absolutely **correct** .
REF:	**Where** were you when you **began** to **feel** bad ?
BERT:	Where **were** you when you began to feel **bad** ?
REF:	He is taller than the Indian , **not** so **tall** as **Gilchrist** .
BERT:	He is **taller** than the Indian , **not** so tall as **Gilchrist** .

Table 5: Typical 3-way prominence predictions of BERT compared to reference labels.

Given that inter-speaker agreement on pitch accent placement is somewhere between 80 and 90% (Yuan et al., 2005), we cannot expect large improvements without speaker-specific modelling.

The source speech data contains multitude of genres ranging from non-fiction to metric poems with fixed prominence patterns and children's stories with high proportion of words emphasized. The difference in genres could impact the test results. Moreover, the books included in the source speech data are all from pre-1923, whereas BERT and GloVe are pre-trained with contemporary texts. We expect that the difference between BERT and other models would be higher with a dataset drawn from a more contemporary source. As noted in Section 3, the difference between BERT and BiLSTM is much bigger with the The Boston University radio news corpus test set (+3.9% compared to +1.1% with our test set). This could be due to the genre, with The Boston University radio news corpus being derived from a more contemporary source.

Overall, our results for BERT highlight the importance of pre-training of the word representations. As we noticed, already with as little as 10% of the training data, BERT outperforms the other models when they are trained on the entire training set. This suggests that BERT has implicitly learned syntactic or semantic information relevant for the prosody prediction task. Our results are in line with the earlier results by Stehwien et al. (2018) and Rendel et al. (2016) who showed that pre-trained word embeddings improve model performance in the prominence prediction task. Table 5 lists five randomly selected examples from the test set and shows the prominence predictions by BERT compared to the reference annotation. These examples indicate that even if the overall accuracy of the model is not high, the predictions still look plausible in isolation.

Finally, the classifiers in this paper are trained on single sentences, losing any discourse-level information and relations to surrounding context. Increasing the context to contain, e.g., also previous sentences could improve the results.

7 Conclusion

In this paper we have introduced a new NLP dataset and benchmark for predicting prosodic prominence from text, which to our knowledge is the largest publicly available dataset with prosodic labels. We described the dataset creation and the resulting benchmark and showed that various sequence labeling methods can be applied to the task of predicting prosodic prominence using the dataset.

Our experimental results show that BERT outperforms the other models with just up to 10% of the training data, highlighting the effectiveness of pre-training for the task. It also highlights that the implicit syntactic or semantic features BERT has learned during pre-training are relevant for the specific task of predicting prosodic prominence.

We also discussed a number of limitations of the automatic annotation system, as well as our current models. Based on this discussion, and more broadly, on the findings of this paper, we want to focus our future research activities in two fronts. Firstly, we will further develop the dataset annotation pipeline, improving the quality of prominence annotation and adding prosodic boundary labels. Secondly, we will further de-

velop methods and models for improved prediction of prosodic prominence. In particular, as our results have shown that pre-training helps in the task, fine-tuning BERT with data involving features that are known to impact prosodic prominence (like part-of-speech tagged data) before training on the prosody dataset could help to improve the model performance. Furthermore, we will look at speaker-aware models, genre adaptation, and models for increased context. And, finally, our ultimate goal is to incorporate these methods into the development of a state-of-the-art text-to-speech synthesizer.

Acknowledgments

Talman, Celikkanat and Tiedemann are supported by the FoTran project, funded by the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme (grant agreement no. 771113).

We also gratefully acknowledges the support of the Academy of Finland through projects no. 314062 from the ICT 2023 call on Computation, Machine Learning and Artificial Intelligence, no. 1293348 from the call on Digital Humanities, and an Academy Fellowship project no. 309575.

References

Dwight Bolinger. 1972. Accent is predictable (if you're a mind-reader). *Language*, pages 633–644.

Joan W Bresnan. 1973. Sentence stress and syntactic transformations. In *Approaches to natural language*, pages 3–47. Springer.

Noam Chomsky and Morris Halle. 1968. The sound pattern of english.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Michelle L Gregory and Yasemin Altun. 2004. Using conditional random fields to predict pitch accents in conversational speech. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL-2004)*, page 677. Association for Computational Linguistics.

Julia Hirschberg. 1993. Pitch accent in context predicting intonational prominence from text. *Artificial Intelligence*, 63(1-2):305–340.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Daniel Jurafsky. 1996. A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive science*, 20(2):137–194.

Daniel Jurafsky, Alan Bell, Michelle Gregory, and William D Raymond. 2001. Probabilistic relations between words: Evidence from reduction in lexical production. *Typological studies in language*, 45:229–254.

Sofoklis Kakouros, Joris Pelemans, Lyan Verwimp, Patrick Wambacq, and Okko Räsänen. 2016. Analyzing the contribution of top-down lexical and bottom-up acoustic cues in the detection of sentence prominence. In *INTERSPEECH*, pages 1074–1078.

Sofoklis Kakouros and Okko Räsänen. 2016. 3pro–an unsupervised method for the automatic detection of sentence prominence in speech. *Speech Communication*, 82:67–84.

Gina-Anne Levow. 2008. Automatic prosodic labeling with conditional random fields and rich acoustic features. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*.

Philip Lieberman. 1960. Some acoustic correlates of word stress in american english. *The Journal of the Acoustical Society of America*, 32(4):451–454.

Erwin Marsi, Martin Reynaert, Antal van den Bosch, Walter Daelemans, and Veronique Hoste. 2003. Learning to predict pitch accents and prosodic boundaries in dutch. In *41st Annual meeting of the Association for Computational Linguistics : proceedings of the conference*, pages 489–496. Association for Computational Linguistics.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, pages 498–502.

Tomas Mikolov, Kai Chen, Greg S. Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of ICLR*.

Yoonsook Mo, Jennifer Cole, and Eun-Kyung Lee. 2008. Naïve listeners prominence and boundary perception. *Proc. Speech Prosody, Campinas, Brazil*, pages 735–738.

Thomas Mueller, Helmut Schmid, and Hinrich Schütze. 2013. Efficient higher-order CRFs for morphological tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language*

Processing, pages 322–332. Association for Computational Linguistics.

Ani Nenkova, Jason Brenier, Anubha Kothari, Sasha Calhoun, Laura Whitton, David Beaver, and Dan Jurafsky. 2007. To memorize or to predict: Prominence labeling in conversational speech. In *Proceedings of the Human Language Technology Conference of the North American chapter of the Association for Computational Linguistics (NAACL-HLT-2007)*, pages 9–16.

Mari Ostendorf, Patti J Price, and Stefanie Shattuck-Hufnagel. 1995. The Boston University radio news corpus. *Linguistic Data Consortium*, pages 1–19.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. LibriSpeech: an ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. IEEE.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics.

Asaf Rendel, Raul Fernandez, Ron Hoory, and Bhuvana Ramabhadran. 2016. Using continuous lexical embeddings to improve symbolic-prosody prediction in a text-to-speech front-end. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5655–5659. IEEE.

Andrew Rosenberg. 2009. *Automatic detection and classification of prosodic events*. Columbia University.

Sabrina Stehwien, Ngoc Thang Vu, and Antje Schweitzer. 2018. Effects of word embeddings on neural network-based pitch accent detection. In *9th International Conference on Speech Prosody*, pages 719–723.

Karl Stratos and Michael Collins. 2015. Simple semi-supervised POS tagging. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*.

Antti Suni, Juraj Šimko, Daniel Aalto, and Martti Vainio. 2017. Hierarchical representation and estimation of prosody using continuous wavelet transform. *Computer Speech & Language*, 45:123–136.

Jacques Terken and Dik Hermes. 2000. The perception of prosodic prominence. In *Prosody: Theory and experiment*, pages 89–127. Springer.

Michael Wagner and Duane G Watson. 2010. Experimental and theoretical advances in prosody: A review. *Language and cognitive processes*, 25(7-9):905–945.

Dagen Wang and Shrikanth Narayanan. 2007. An acoustic measure for word prominence in spontaneous speech. *IEEE transactions on audio, speech, and language processing*, 15(2):690–701.

Oliver Watts. 2012. *Unsupervised Learning for Text-to-Speech Synthesis*. Ph.D. thesis, University of Edinburgh.

Tae-Jin Yoon, Sandra Chavarria, Jennifer Cole, and Mark Hasegawa-Johnson. 2004. Intertranscriber reliability of prosodic labeling on telephone conversation using tobi. In *Eighth International Conference on Spoken Language Processing*.

Jiahong Yuan, Jason M Brenier, and Daniel Jurafsky. 2005. Pitch accent prediction: Effects of genre and speaker. In *Ninth European Conference on Speech Communication and Technology*.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. LibriTTS: A corpus derived from LibriSpeech for text-to-speech. *arXiv preprint arXiv:1904.02882*.