HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

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

Aarne Talman, Antti Suni, Hande Celikkanat, Sofoklis Kakouros, Jörg Tiedemann and Martti Vainio

October 2, 2019

Department of Digital Humanities

University of Helsinki

Outline

Outline

Introduction: Prosody

Prosodic prominence: the amount of emphasis that a speaker gives to a word.

- Prosody has been widely studied in phonetics and speech processing.
- Research on text-based natural language processing (NLP) methods is somewhat limited, even in text-to-speech synthesis domain.
- The main reason is lack of suitable and large enough datasets for the modern data-hungry approaches.

Predicting prosodic prominence from text:

- Given text, the task of predicting the prominence of each word in a sentence either as a continuous value or a discrete value.

Research question:

- Can we use text-based NLP methods to predict speech prosody from text?

Outline

Helsinki Prosody Corpus

We introduce a new NLP benchmark and the largest annotated dataset for predicting prosodic prominence from text, with automatically generated high-quality annotations for the recently published LibriTTS corpus (Zen et al., 2019).

- For annotation we used the Wavelet Prosody Analyzer toolkit[1] which implements the method described in Suni et al. (2017).
    1. 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.
    4. Performing a continuous wavelet transform (CWT) on the composite signal and extracting continuous prominence values as lines of maximum amplitude across wavelet scales.
- The method assumes that the louder, the longer, and the higher the acoustic signal for a word is, the more prominent it is.

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

Continuous Wavelet Transform Annotation Method

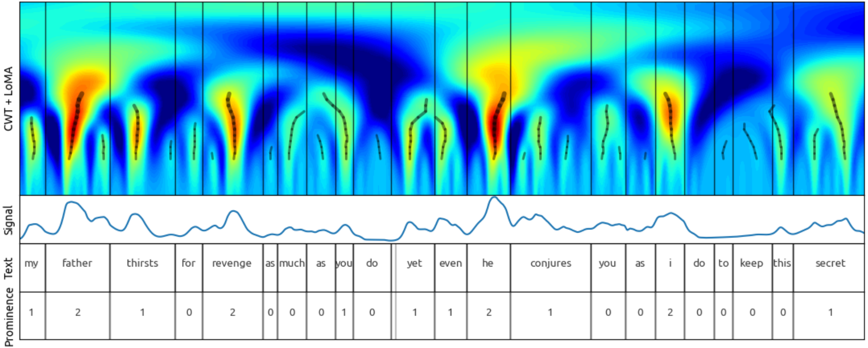

Figure 1: Continuous Wavelet Transform Annotation method.

Helsinki Prosody Corpus: Statistics

The resulting dataset contains over 2.8 Million annotated tokens of English text divided into two training sets, a dev set and a test set.

				non-prominent	prominent	
sets (clean)	**speakers**	**sentences**	**tokens**	**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

Figure 2: Dataset statistics

Helsinki Prosody Corpus

`https://github.com/Helsinki-NLP/prosody`

- Text files with one token per line.
- Sentences separated with a line: `<file> file_name.txt`, referring to the source file in LibriTTS.
- Each line has five items separated with tabs (with NA for punctuation):
  1. Token
  2. Discrete prominence label: 0 (non-prominent), 1 (prominent), 2 (highly prominent)
  3. Discrete word boundary label: 0, 1, 2
  4. Continuous prominence value
  5. Continuous word boundary value

- Example sentence:

```
<file>    6829_68769_000053_000002.txt
That's    1    1    0.984    0.842
how       2    0    2.122    0.000
all       1    1    0.463    1.411
the       0    0    0.009    0.432
trouble   2    1    1.549    0.634
came      0    0    0.144    0.097
about     1    2    0.948    2.0
.         NA   NA   NA       NA
```

The new dataset allows us to treat prosody prediction as a text-based sequence labeling task, like PoS tagging or NER.

Outline

Models

We performed experiments from with multiple feature-based and neural models.

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

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

For pre-trained BERT we used the Huggingface Transformers library.

Results

Experimental results for different models trained on the `train-360` dataset.

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%

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

- 3-way classification task uses all the labels 0, 1 and 2
- 2-way classification task combines 1 and 2

Results: Confusion matrices

3-way classification task confusion matrices for the BERT and BiLSTM models.

		Predicted 0	Predicted 1	Predicted 2	*recall*
Gold	0	35567	5602	2043	**82.3%**
Gold	1	5943	11589	6987	**47.3%**
Gold	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	Predicted 1	Predicted 2	*recall*
Gold	0	35321	6157	1734	**81.0%**
Gold	1	6221	12275	6019	**46.4%**
Gold	2	2058	8014	12172	**61.1%**
	precision	**81.7%**	**50.1%**	**54.7%**	

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

Results: Learning curves

BERT 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 case.

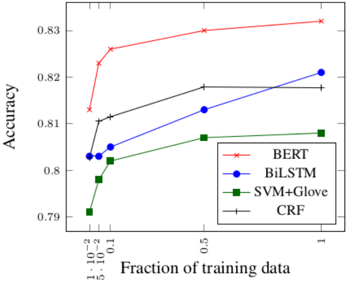

Figure 6: Test accuracy with different size subsets of the training data for the 2-way classification task.

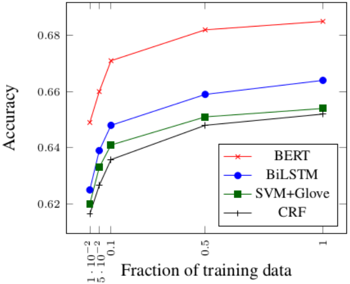

Figure 7: Test accuracy with different size subsets of the training data for the 3-way classification task.

Outline

Discussion

Although our experiments show that prosodic prominence can reasonably well be predicted from text, the scores are still quite low.

Some reasons could be:

- Errors in automatic alignment, signal processing, and quantization introduce noise to the labels.
- Different speakers have different accents, varying reading proficiency, and reading tempo, which all impact the consistency of the labeling.
- The source speech data contains genres ranging from non-fiction to metric poems with fixed prominence patterns and children's stories. The difference in genres could impact the test results.
- The books included in the source speech data are all from pre-1923, whereas BERT and GloVe are pre-trained with contemporary texts.

Conclusion

- We introduce a new NLP benchmark for predicting prosodic prominence from text.
- We publish the largest publicly available datasets with prosodic labels.
- The new dataset allows us to treat prosody prediction as a normal sequence labeling task and apply text-based models to the task.
- We test wide variaty of models and show that BERT outperforms the other approaches even with a very small subset of the training data.

Data and code:

https://github.com/Helsinki-NLP/prosody

aarne.talman@helsinki.fi

FoTRAN Found in Translation
European Research Council
Established by the European Commission
ACADEMY OF FINLAND