# Online Appendix for "Detecting Local Insights from Global Labels: Supervised & Zero-Shot Sequence Labeling via a Convolutional Decomposition"

Allen Schmaltz
Department of Epidemiology
Harvard University
aschmaltz@hsph.harvard.edu

*Additional results.*

## 1. Introduction

We provide comparisons to the more recent work of Bujel, Yannakoudakis, and Rei (2021), which considers weighted attention over Transformer models for zero-shot binary sequence labeling. In this context, we also provide a correspondence to previous models on data similar to the CoNLL 2010 task considered in earlier works.

This provides additional evidence that the inductive bias of the proposed method is particularly conducive to this type of class-conditional feature detection. Across these additional datasets, we find that our proposed approach for zero-shot sequence labeling is at least as effective—and often significantly more so—than alternatives, while also enabling the additional properties described in the main text.

In Section E we consider the task of grammatical error detection with the FCE dataset, as used in Section 4 of the main text; in Section F we compare against the BEA 2019 grammatical error detection dataset; and in Section G we report results on publicly available data similar to the original CoNLL 2010 task.

## Appendix E: Grammatical Error Detection: Additional Results

### E.1 Data

We use the same FCE data of Section 4 of the paper, evaluating on the FCE test set.

### E.2 Models

We use a model identical to UNICNN+BERT of Section 4 of the paper, with the one difference that we use a pre-trained $\text{BERT}_{\text{BASE}}$ Transformer since Bujel, Yannakoudakis, and Rei (2021) uses a Transformer with a $\text{BERT}_{\text{BASE}}$ architecture, RoBERTa (Liu et al. 2019). We use the label UNICNN+$\text{BERT}_{\text{BASE}}$ for this model.

Table E.1: Additional FCE zero-shot sequence labeling test set results (cf., Table 1 of the main text). Models marked with † indicate results stated in their respective papers. With the exception of LIME, all models only have access to sentence-level labels while training. The sentence-level $F_1$ scores for the CNN models are from the fully-connected layer and are provided for reference. Token-level evaluation is the same across papers, as further indicated by a similar RANDOM baseline from Bujel, Yannakoudakis, and Rei (2021).

| | Sent | Token-level | | | |
|---|---|---|---|---|---|
| Model | $F_1$ | P | R | $F_1$ | $F_{0.5}$ |
| RANDOM† | - | 15.11 | 49.81 | 23.19 | 17.56 |
| RANDOM | 58.30 | 15.30 | 50.07 | 23.44 | 17.79 |
| MAJORITYCLASS | 80.88 | 15.20 | 100. | 26.39 | 18.31 |
| LIME† | 84.51 | 19.06 | 34.70 | 24.60 | 20.95 |
| LSTM-ATTN-SW† | 85.14 | 28.04 | 29.91 | 28.27 | 28.40 |
| ROBERTA$_{\text{BASE}}$+WSA† | 85.62 | 20.76 | 85.36 | 33.31 | 24.46 |
| UNICNN$_{M=2}$+BERT$_{\text{BASE}}$ | 86.22 | 57.91 | 19.33 | 28.99 | 41.39 |
| UNICNN+BERT$_{\text{BASE}}$ | 86.29 | 53.17 | 35.37 | 42.48 | 48.31 |

*Reference Model.* For instructive purposes, we also train a limited capacity[1] version of the model with only two filters, $M = 2$, UNICNN$_{M=2}$+BERT$_{\text{BASE}}$.

*Previous Models.* The recent work of Bujel, Yannakoudakis, and Rei (2021) adapts the soft attention-based approach used for LSTMs of Rei and Søgaard (2018) to multi-headed Transformers, finding that a weighted variant, for which we use the label ROBERTA$_{\text{BASE}}$+WSA, yielded higher $F_1$ scores and qualitatively sharper detections than the unweighted version. In contrast, using scores from the multi-headed attention directly required setting a threshold based on held-out token-level labels, and even then, resulted in very diffuse detections only marginally better than a random baseline. We also include the reported results for LIME (Ribeiro, Singh, and Guestrin 2016) under this RoBERTa$_{\text{BASE}}$ model, noting that the LIME baseline, is not truly "zero-shot" sequence labeling since the threshold is learned based on token-level labels. Finally, we use the label LSTM-ATTN-SW for the model of Rei and Søgaard (2018), as in the main text. We include the $F_1$ scores stated in the earlier work, and we also calculate $F_{0.5}$ scores based on the reported recall and precision results.

### E.3 FCE Additional Results

Table E.1 contains the additional baseline results. As expected based on the previously observed quantitative and qualitative results, it is challenging to achieve similar token-level detection results using soft-attention approaches over the Transformer. In fact, the

1 Due to the max-pooling operation, $M$ is in effect a hard upper-bound on the number of tokens in a sentence that can have non-zero token-level predictions ($s_n^{+-}$) using this approach.

results are substantively lower, even though the sentence-level $F_1$ scores are the same for all practical purposes.[2] Additionally, a LIME baseline does not correlate particularly well with the human-annotated labels.

The kernel-width-one CNN and linear layer are able to bottleneck the signal from the deep network in a manner corresponding roughly to the token-level labels in these datasets. As we see with UNICNN$_{M=2}$+BERT$_{\text{BASE}}$, two filters are sufficient for achieving relatively high precision with similar sentence-level effectiveness. Capacity can then be increased by simply increasing the number of filters, $M$, which increases recall. (Separately, this also yields a representative vector for each token, as described in the main text.) In contrast, increasing soft-attention capacity as with multi-headed attention, while useful—and perhaps critical—in lower layers of the Transformer, leads to very diffuse detections in the final layer vis-a-vis human-annotated token-level labels in these datasets.

## Appendix F: Grammatical Error Detection: BEA 2019

### F.4 Data

We use the data of the BEA-2019 Shared Task on Grammatical Error Correction (Bryant et al. 2019) as an additional grammatical error detection dataset. The task is the same as that with the FCE dataset used in the main text, but the BEA-2019 data is reported to include sentences across a greater diversity of language proficiency. We use the split indexes provided by Bujel, Yannakoudakis, and Rei (2021), using 10% of the training set for the dev set and the original Shared Task dev set as the held-out test.

### F.5 Models

We report results for the same main models as in Section E, noting that the LSTM-ATTN-SW result is that reported in Bujel, Yannakoudakis, and Rei (2021). We additionally fine-tune with the min-max loss for reference on this new dataset, UNICNN+BERT$_{\text{BASE}}$+MM.

### F.6 BEA 2019 Results

Table F.1 shows that the overall patterns are similar to those on the FCE dataset. The BEA 2019 dataset appears to be more challenging, perhaps owing to the greater diversity of writers, despite having a similar training set size as the FCE data.[3] As with the FCE data, we see that our approach yields a significantly stronger sequence labeler than the alternatives.

## Appendix G: Uncertainty Tag Detection: CoNLL 2010

Previously reported results on the CoNLL 2010 Shared Task (Farkas et al. 2010) data suggest a significantly easier zero-shot sequence labeling task than the grammar tasks.

2 The difference is also not explained by a smaller Transformer. In fact, on this dataset, the BERT$_{\text{BASE}}$ variation is no worse than the BERT$_{\text{LARGE}}$ version used in the main text.

3 The differences in the distribution are also evident by the lower RANDOM and MAJORITYCLASS baselines compared to the FCE data.

Table F.1: BEA 2019 zero-shot sequence labeling test set results. Models marked with † indicate results stated in existing works. With the exception of LIME, all models only have access to sentence-level labels while training. The sentence-level $F_1$ scores for the CNN models are from the fully-connected layer and are provided for reference. Token-level evaluation is the same across papers, as further indicated by a similar RANDOM baseline from Bujel, Yannakoudakis, and Rei (2021).

| | Sent | Token-level | | | |
|---|---|---|---|---|---|
| Model | $F_1$ | P | R | $F_1$ | $F_{0.5}$ |
| RANDOM† | - | 10.05 | 50.00 | 16.73 | 11.96 |
| RANDOM | 57.13 | 10.08 | 50.02 | 16.78 | 12.00 |
| MAJORITYCLASS | 78.90 | 10.11 | 100. | 18.36 | 12.32 |
| LIME† | 83.66 | 13.49 | 1.13 | 2.09 | 4.23 |
| LSTM-ATTN-SW† | 81.27 | 10.93 | 61.63 | 18.53 | 13.08 |
| ROBERTA$_{\text{BASE}}$+WSA† | 83.68 | 14.20 | 85.49 | 24.35 | 17.04 |
| UNICNN+BERT$_{\text{BASE}}$ | 84.49 | 37.26 | 37.61 | 37.43 | 37.33 |
| UNICNN+BERT$_{\text{BASE}}$+MM | 84.20 | 45.18 | 27.79 | 34.42 | 40.16 |

At the same time, the $F_1$ scores, and especially the $F_{0.5}$ scores, of more recent Transformer approaches fall below those of soft-attention over LSTMs. We investigate this data distribution further in this section with our model.

### G.7 Data

The original CoNLL 2010 Shared Task data was not publicly available, so we instead reprocess the publicly available Szeged Uncertainty Corpus.[4] This is ostensibly the same training data as the original Shared Task, but the held-out test split is different. We provide our data processing scripts for future replications. We split the data randomly by documents, not sentences, to avoid document overlap across splits, and we remove any sentence overlap between the test split and training and dev. This results in 16,198 sentences for training, 1,960 sentences for dev, and 1,940 sentences for test. The training set is about half the size of that of the grammar sets.

We assign positive token labels ($y_n = 1$) to any token contained within a `ccue` XML tag, and any sentence with at least one positive token receives a positive sentence-level label ($Y = 1$). These tags correspond to "uncertainty" cues, to which we defer to the original reference for further description. Here we are less interested in the semantic meaning of the tags, and more interested in their distribution compared to the labels of the grammar tasks. The tags are very rare relative to the total number of tokens, with only around 1% of tokens in the test set having positive labels, but occur with sufficient lexical and contextual regularity to be nonetheless relatively easy for the models to detect, with some exceptions discussed below.

4 This data is publicly available at https://rgai.inf.u-szeged.hu/file/139 and described further at https://rgai.inf.u-szeged.hu/node/160.

Table G.1: CoNLL 2010 zero-shot sequence labeling test set results. Note that this test split differs from that of the original Shared Task. With the exception of UNICNN+BERT$_{\text{BASE}_{\text{UNCASED}}}$+S*, all models only have access to sentence-level labels while training.

| | Sent | Token-level | | | |
|---|---|---|---|---|---|
| Model | $F_1$ | P | R | $F_1$ | $F_{0.5}$ |
| RANDOM | 31.1 | 1.30 | 52.28 | 2.53 | 1.61 |
| MAJORITYCLASS | 35.57 | 1.24 | 100. | 2.45 | 1.55 |
| LSTM-ATTN-SW | 89.18 | 87.5 | 73.43 | 79.85 | 84.27 |
| ROBERTA$_{\text{BASE}}$+WSA | 89.97 | 27.65 | 91.03 | 42.41 | 32.12 |
| UNICNN+BERT$_{\text{BASE}_{\text{UNCASED}}}$ | 88.08 | 42.74 | 84.60 | 56.79 | 47.43 |
| UNICNN+BERT$_{\text{BASE}_{\text{UNCASED}}}$+MM | 87.45 | 86.69 | 70.56 | 77.80 | 82.90 |
| UNICNN+BERT$_{\text{BASE}_{\text{UNCASED}}}$+MM+EXAG | - | 90.91 | 65.99 | 76.47 | 84.53 |
| UNICNN+BERT$_{\text{BASE}_{\text{UNCASED}}}$+MM+K$_8$NN$_{\text{DIST.}}$ | - | 85.4 | 72.25 | 78.28 | 82.40 |
| UNICNN+BERT$_{\text{BASE}_{\text{UNCASED}}}$+S* | 89.05 | 90.73 | 76.14 | 82.80 | 87.38 |

### G.8 Models

Given the new splits, we re-train the ROBERTA$_{\text{BASE}}$+WSA model from Bujel, Yannakoudakis, and Rei (2021) using the publicly available code and configuration for the original CoNLL Shared Task. We similarly re-train the LSTM-ATTN-SW model from Rei and Søgaard (2018), which has been reported to out-perform more recent Transformer approaches on this dataset, in contrast to results on the grammar datasets. We lowercase and tokenize the data as done in earlier work. Our base model, UNICNN+BERT$_{\text{BASE}_{\text{UNCASED}}}$, uses the uncased smaller BERT model due to the aforementioned lowercasing and for comparison to the earlier Transformer work. We fine-tune 300 dimensional Glove embeddings, as with LSTM-ATTN-SW. We fine-tune the model with the min-max loss, UNICNN+BERT$_{\text{BASE}_{\text{UNCASED}}}$+MM, for which we also consider the EXAG inference-time decision rule, UNICNN+BERT$_{\text{BASE}_{\text{UNCASED}}}$+MM+EXAG, and a distance-weighted K-NN approximation, UNICNN+BERT$_{\text{BASE}_{\text{UNCASED}}}$+MM+K$_8$NN$_{\text{DIST.}}$. Finally, to provide a rough empirical upper-bound on the zero-shot sequence labeling effectiveness, we also train a fully-supervised model, UNICNN+BERT$_{\text{BASE}_{\text{UNCASED}}}$+S*, by fine-tuning the base model with token-level labels.

### G.9 CoNLL 2010 Results

The results on the test set appear in Table G.1. Overall, this test split is slightly less challenging than the original held-out test, which annotated additional held-out articles, but the overall pattern of LSTM-ATTN-SW outperforming the soft-attention variation over a Transformer, ROBERTA$_{\text{BASE}}$+WSA, despite similar sentence-level scores, is as previously reported. More diffuse—and higher recall—predictions were also observed on the grammar sets with the ROBERTA$_{\text{BASE}}$+WSA model, but the impact here is particularly exaggerated in the $F$ scores due to the sparsity of the ground-truth labels in this dataset.

The overall scores tend to be much higher than those for the grammar datasets, despite the extreme sparsity of the labeled tokens vs. the total number of tokens. In fact, the most effective models approach the fully-supervised model UNICNN+BERT$_{\text{BASE}_{\text{UNCASED}}}$+S*. As on the other datasets, the inference-time decision rule +EXAG improves precision, and the K-NN approximation is at least as effective as the corresponding original model. The UNICNN+BERT$_{\text{BASE}_{\text{UNCASED}}}$+MM model, which imposes the min-max constraint, closes the gap with the min-max LSTM model, while the UNICNN+BERT$_{\text{BASE}_{\text{UNCASED}}}$ model results in more diffuse predictions, even though the sentence-level $F_1$ scores are both around 88.[5] This difference is readily evident by simply visualizing the detections; in practice, for new domains or datasets, both models can be trained and compared to better understand the data and suitability of the min-max, or related, constraints. We show examples in Table G.2. At this level, the differences between the most effective models are likely not practically significant.

---

In the most general case, without additional assumptions, determining token-level labels from document-level labels is an under-defined task. Multiple label annotation schemes could be consistent with the document-level labels, which is an independent challenge of the parameters of the neural networks themselves being non-identifiable. Despite these intrinsic challenges, we have proposed and analyzed an approach that is likely to be useful in many settings in practice. The zero-shot sequence labeling approach we have proposed is consistently at least as effective as alternatives as we have identified an inductive bias over the deep networks that corresponds to the annotated labels across the observed datasets at least as closely as known alternatives. In this way, we can leverage the density estimation of a deep network, pre-trained over large amounts of data, for class-conditional feature detection. Combined with the additional approaches linking the predictions to a support set with known labels, we can proactively leverage the deep networks to analyze datasets and models at lower resolutions of the input than that of the available training labels.

5 For the UNICNN+BERT$_{\text{BASE}_{\text{UNCASED}}}$+MM and UNICNN+BERT$_{\text{BASE}_{\text{UNCASED}}}$+S* models we take as the reference sentence-level prediction the max token-level contribution in each sentence, $\hat{Y} = \text{sgn}(s^{+-}_{max})$, rather than the softmax output from the fully-connected layer. This is based on the document-level $F_1$ scores on the dev set. As with the experiments in the main text, we do not impose a global constraint on the final layer when fine-tuning the min-max and supervised losses, so this alternative can be useful if the final layer's parameters change significantly during fine-tuning. These two options for document-level classification are sufficient for our observed binary datasets, but a global constraint can be useful in some cases when extending to multi-class and multi-label settings, which we leave for future work.

Table G.2: Two example sentences from the new CoNLL 2010 test set, across the zero-shot sequence labeling models. Positive predictions are underlined, with true positive predictions in blue and false positive predictions in red. The ground-truth labeled sentence is marked TRUE, with ground-truth token-level labels underlined.

| | |
|---|---|
| **Sentence 779** | |
| TRUE | The BCL6 gene encodes a 95-kDa protein containing six C-terminal zinc-finger motifs and an N-terminal POZ domain, suggesting that it may function as a transcription factor. |
| LSTM-ATTN-SW | The BCL6 gene encodes a 95-kDa protein containing six C-terminal zinc-finger motifs and an N-terminal POZ domain, suggesting that it may function as a transcription factor. |
| ROBERTA$_{\text{BASE}}$+WSA | The BCL6 gene encodes a 95-kDa protein containing six C-terminal zinc-finger motifs and an N-terminal POZ domain, suggesting that it may function as a transcription factor. |
| UNICNN+BERT$_{\text{BASE}_{\text{UNCASED}}}$ | The BCL6 gene encodes a 95-kDa protein containing six C-terminal zinc-finger motifs and an N-terminal POZ domain, suggesting that it may function as a transcription factor. |
| UNICNN+BERT$_{\text{BASE}_{\text{UNCASED}}}$+MM | The BCL6 gene encodes a 95-kDa protein containing six C-terminal zinc-finger motifs and an N-terminal POZ domain, suggesting that it may function as a transcription factor. |
| UNICNN+BERT$_{\text{BASE}_{\text{UNCASED}}}$+MM +K$_8$NN$_{\text{DIST.}}$ | The BCL6 gene encodes a 95-kDa protein containing six C-terminal zinc-finger motifs and an N-terminal POZ domain, suggesting that it may function as a transcription factor. |
| **Sentence 1717** | |
| TRUE | However, little is known about the structure-activity relationship and the mechanism by which endotoxin induces Mn SOD. |
| LSTM-ATTN-SW | However, little is known about the structure-activity relationship and the mechanism by which endotoxin induces Mn SOD. |
| ROBERTA$_{\text{BASE}}$+WSA | However, little is known about the structure-activity relationship and the mechanism by which endotoxin induces Mn SOD. |
| UNICNN+BERT$_{\text{BASE}_{\text{UNCASED}}}$ | However, little is known about the structure-activity relationship and the mechanism by which endotoxin induces Mn SOD. |
| UNICNN+BERT$_{\text{BASE}_{\text{UNCASED}}}$+MM | However, little is known about the structure-activity relationship and the mechanism by which endotoxin induces Mn SOD. |
| UNICNN+BERT$_{\text{BASE}_{\text{UNCASED}}}$+MM +K$_8$NN$_{\text{DIST.}}$ | However, little is known about the structure-activity relationship and the mechanism by which endotoxin induces Mn SOD. |

## References


Bryant, Christopher, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. The BEA-2019 Shared Task on Grammatical Error Correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Association for Computational Linguistics, Florence, Italy.

Bujel, Kamil, Helen Yannakoudakis, and Marek Rei. 2021. Zero-shot Sequence Labeling for Transformer-based Sentence Classifiers.

Farkas, Richárd, Veronika Vincze, György Móra, János Csirik, and György Szarvas. 2010. The CoNLL-2010 Shared Task: Learning to Detect Hedges and their Scope in Natural Language Text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning – Shared Task*, pages 1–12, Association for Computational Linguistics, Uppsala, Sweden.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach.

Rei, Marek and Anders Søgaard. 2018. Zero-Shot Sequence Labeling: Transferring Knowledge from Sentences to Tokens. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 293–302, Association for Computational Linguistics, New Orleans, Louisiana.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier.