Lightweight Contextual Logical Structure Recovery

Po-Wei Huang, Abhinav Ramesh Kashyap, Yanxia Qin, Yajing Yang, Min-Yen Kan

National University of Singapore

SDP@COLING 2022
Problem Statement

Classify each individual line into 23 predefined classes that indicate the hierarchy of the document structure.
Can we obtain (near-)SOTA performance on logical structure recovery without relying on feature-rich information, but on context only?
**Dataset**

*Data Source*

- **Original SectLabel Dataset** (*Luong et al., IJDLS 2010*):
  - 20 ACL 2009 Papers
  - 20 CHI 2008 Papers

- **Extended Testing Dataset**:
  - 20 ACL 2020 Papers

- **Unlabelled Dataset**:
  - 570 ACL 2021 Long Papers
  - 1895 NeurIPS 2021 Papers

**8:1:1 Document Split on SectLabel Dataset for Training, Validation and Testing**
Dataset

Category Distribution

Occurrence of Each Category
Contextual Model Construction

Sliding Window Attention

| Author                                | Baseline  | Sliding Window 5 |
|---------------------------------------|-----------|------------------|
| Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. arXiv preprint arXiv:1802.05365, 2018. | reference | reference         |
Pooling Methods for Sentence Embeddings

[CLS] Token

BERT (or any other pretrained transformer model)
Pooling Methods for Sentence Embeddings

**Mean Pooling**

![Diagram showing mean pooling of word embeddings](Image)
Pooling Methods for Sentence Embeddings

Attention Pooling

BERT (or any other pretrained transformer model)
# Semi-Supervised Learning

## Data Augmentation Techniques

| Method                        | Example                  |
|-------------------------------|--------------------------|
| Original                      | Once upon a midnight dreary, while I pondered, weak and weary, |
| Synonym Replacement (EDA)     | **Erstwhile** upon a midnight dreary, while I pondered, weak and weary, |
| Random Insertion (EDA)         | Once upon a midnight dreary, while I pondered, weak and **once** weary, |
| Random Swap (EDA)              | Once upon **I** midnight dreary, while **a** pondered, weak and weary, |
| Random Delete (EDA)            | Once upon a _ dreary, while I pondered, _ and weary, |
| Back Translation              | Once at midnight it was bleak while I was thinking, weak and **tired**, |
Semi-Supervised Learning

Unsupervised Data Augmentation

Diagram:
- Labeled Text
- Unlabeled Text
- Augmented Text
- Backbone Model
- TSA
- Cross-Entropy Loss
- Consistency Loss
Semi-Supervised Learning

FixMatch
Loss Engineering

Training Signal Annealing

Sentence Embedding → Prediction
Sentence Embedding → Prediction
Sentence Embedding → Prediction
Sentence Embedding → Prediction

Training Signal Annealing
Select
Max Prediction
Prob < f(t)

f(t)

Cross Entropy Loss
## Results

| Model                                      | SectLabel | Extended |
|--------------------------------------------|-----------|----------|
|                                            | Macro F1  | Micro F1 | Macro F1  | Micro F1 |
| *SciWING (Ramesh Kashyap and Kan, 2020)*  | 0.732     | 0.900    | -         | -        |
| RoBERTa-Attn Model (OURS)                  | 0.806     | 0.904    | 0.596     | 0.870    |
| RoBERTa-Attn Model + $\text{UDA}_{log}^\dagger$ | 0.784     | 0.906    | **0.669** | **0.887** |
| RoBERTa-Attn Model + $\text{SDA}_{log}^\dagger$ | **0.832** | **0.929** | 0.623     | 0.886    |
| *SectLabel (Luong et al., 2010)*\ddagger  | **0.847** | **0.934** | -         | -        |
Scan to read the full paper!

Connect with the first author!
✉ huangpowei@comp.nus.edu.sg