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

NUS | Computing
National University of Singapore

WING
Web Information Retrieval
Natural Language Processing Group

Lightweight Contextual Logical Structure Recovery

Po-Wei Huang, Abhinav Ramesh Kashyap, Yanxia Qin, Yajing Yang, Min-Yen Kan*
National University of Singapore
{huangpowei,abhinav,qinyx,yang0317,kanmy}@comp.nus.edu.sg

Abstract

Logical structure recovery in scientific
associates text with a semantic section
ticle. Although previous work has disr
the surrounding context of a line, we m
important information by employing li
attention on top of a transformer-base
tific document processing pipeline. V
addition of loss function engineering

ystems (such as Op-
CR) to obtain such
ess cumbersome and
imilar performance
without relying on

by creating a parsi-
t operates on purely
ating such features.

Problem Statement

Classify each individual line into 23 predefined classes that indicate the hierarchy of the document structure.

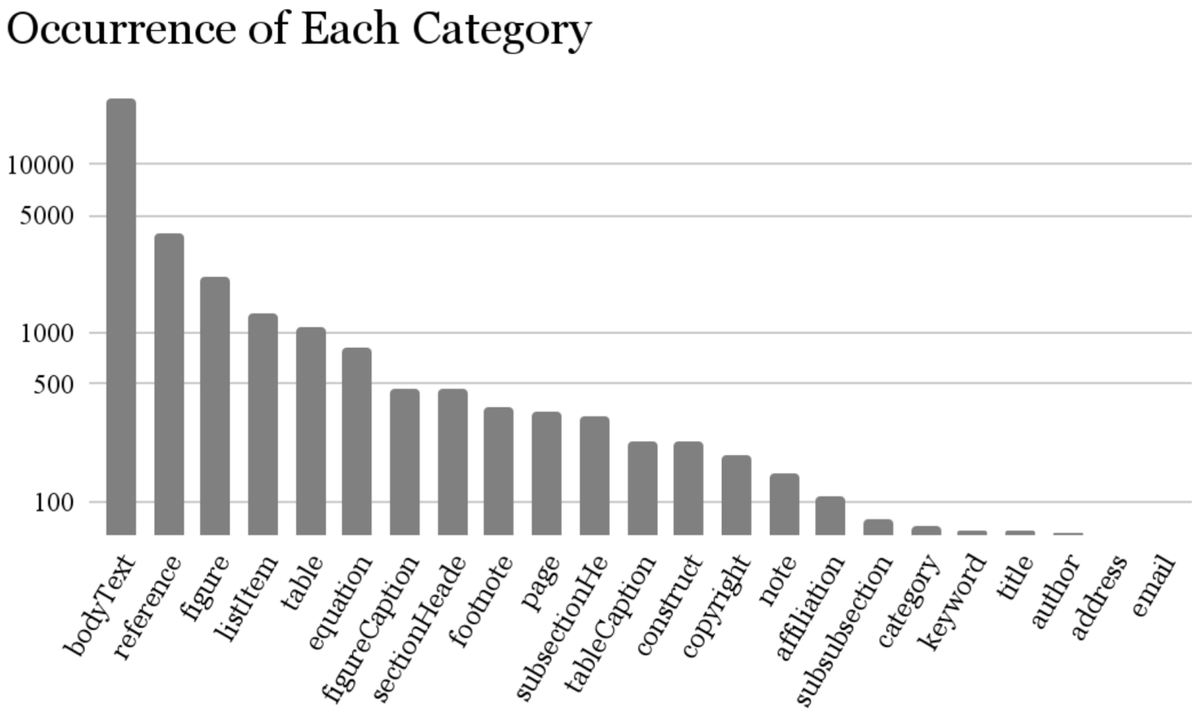
Can we obtain (near-)SOTA performance on logical structure recovery without relying on feature-rich information, but on context only?

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

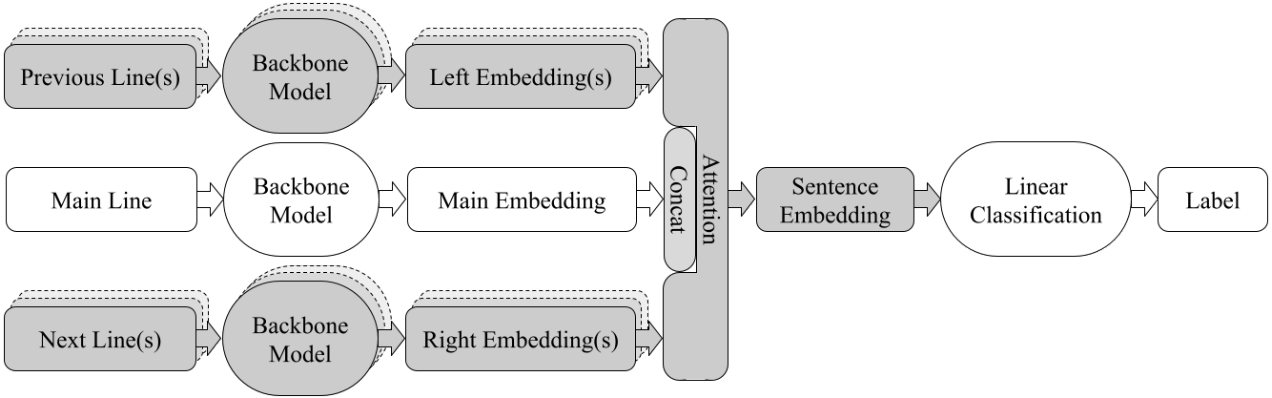
8:1:1 Document Split on SectLabel Dataset for Training, Validation and Testing

Dataset
Category Distribution



Contextual Model Construction

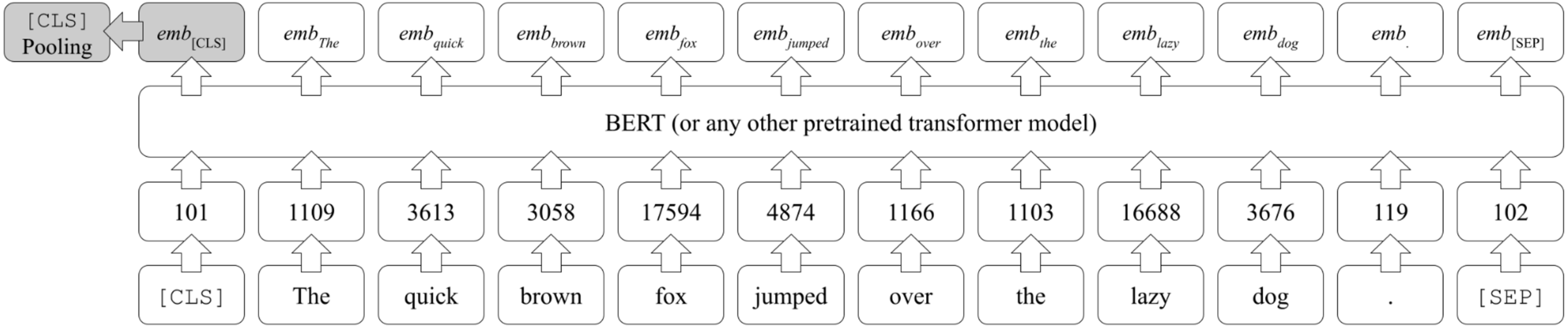
Sliding Window Attention



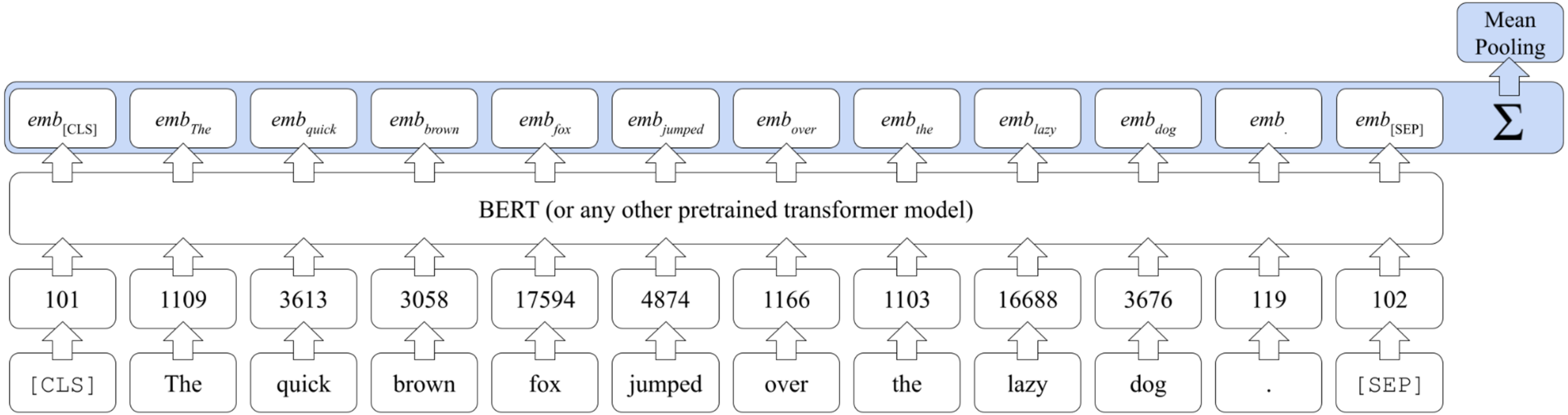
	Baseline	Sliding Window 5
Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt	*author*	reference
Gardner, Christopher Clark, Kenton Lee, and Luke	reference	reference
Zettlemoyer. Deep contextualized word representa-	*bodyText*	reference
tions. arXiv preprint arXiv:1802.05365, 2018.	reference	reference

Pooling Methods for Sentence Embeddings

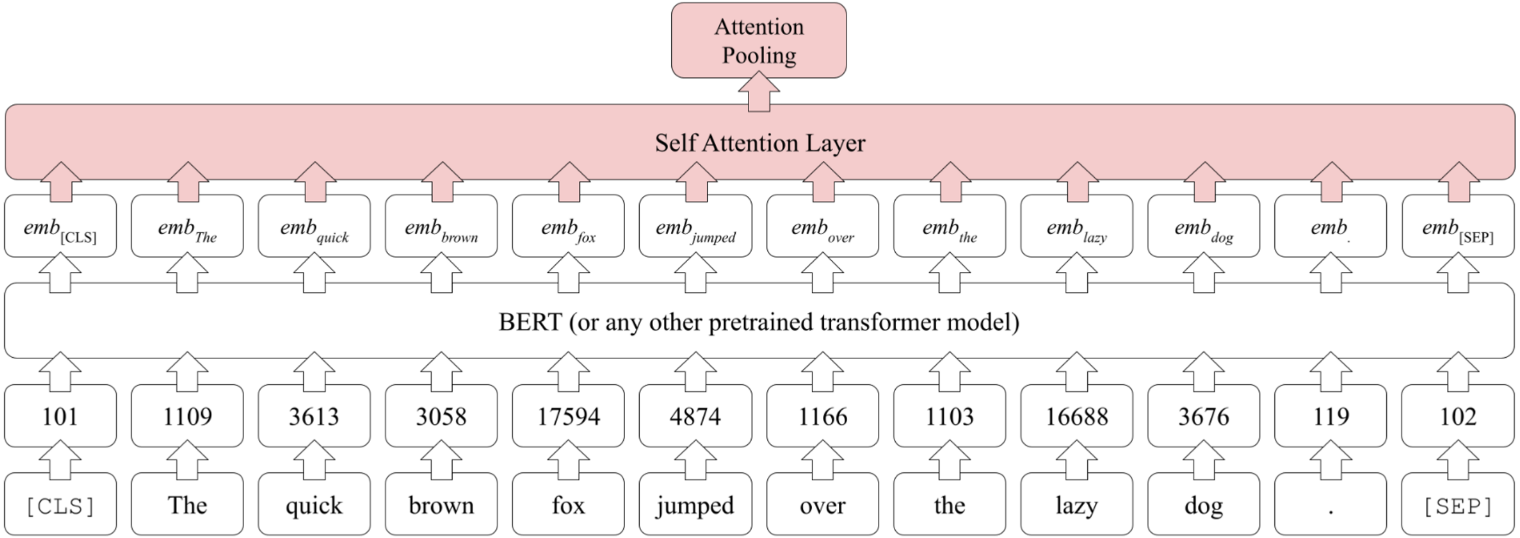
[CLS] *Token*



Pooling Methods for Sentence Embeddings
Mean Pooling



Pooling Methods for Sentence Embeddings
Attention Pooling

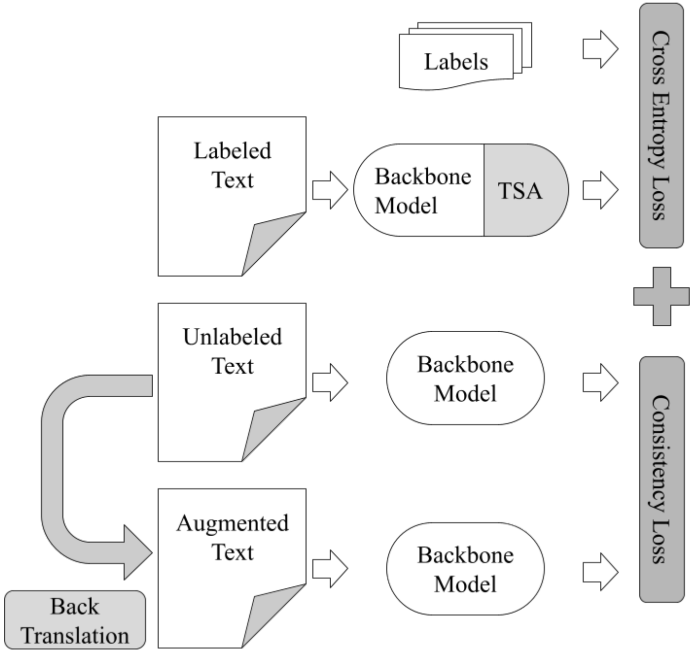


Semi-Supervised Learning

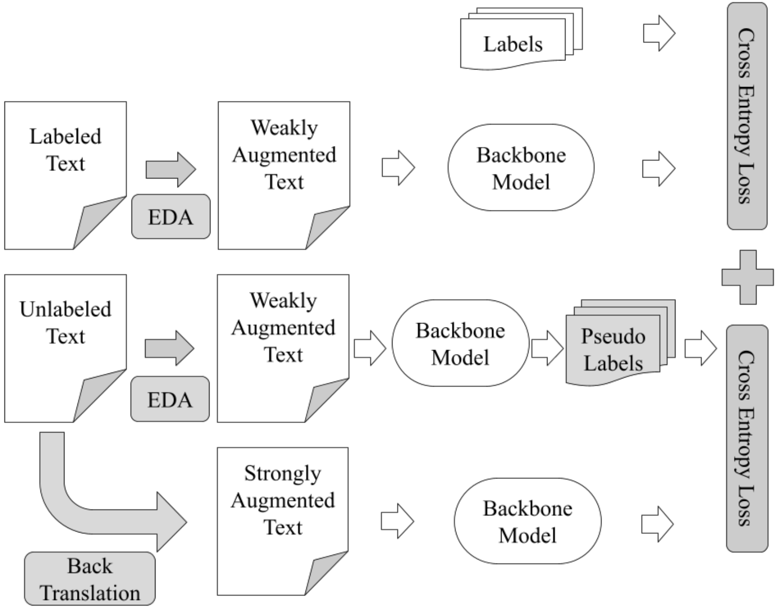
Data Augmentation Techniques

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,

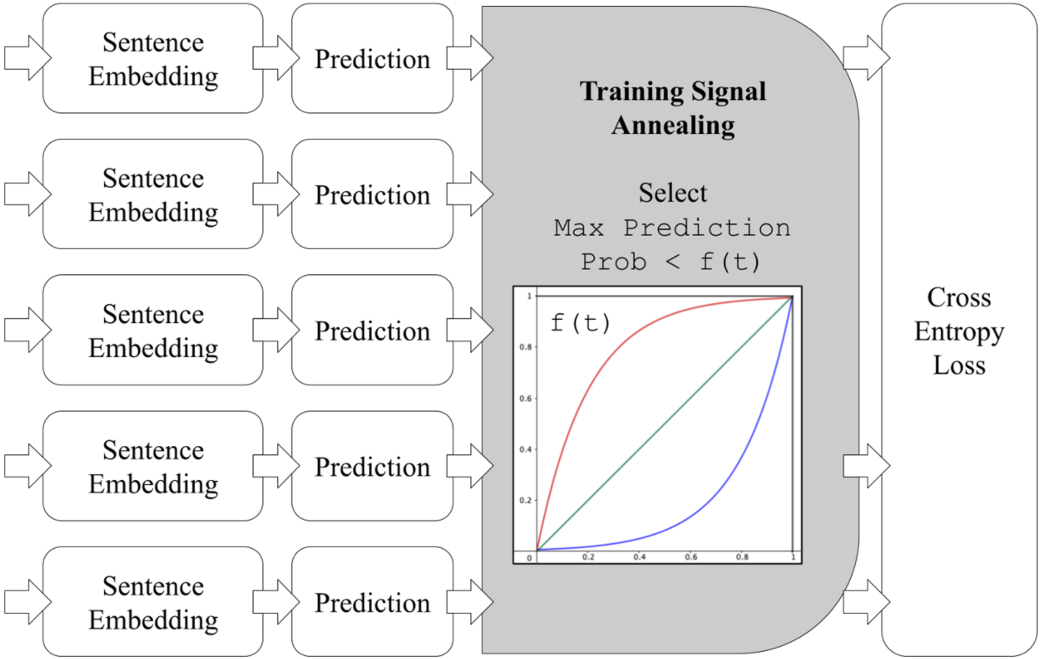
Semi-Supervised Learning
Unsupervised Data Augmentation



Semi-Supervised Learning
FixMatch

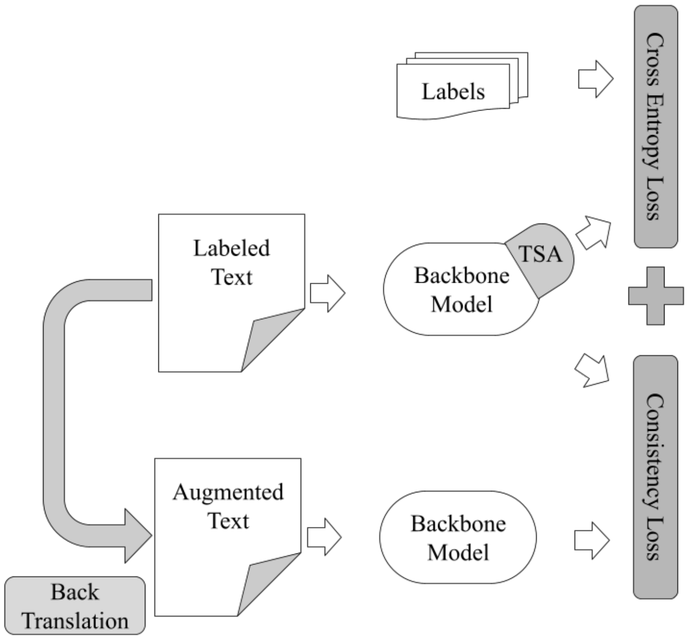


Loss Engineering
Training Signal Annealing



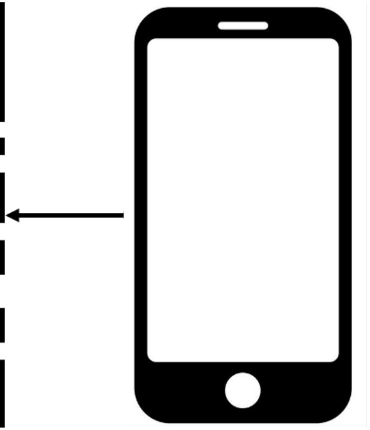
Loss Engineering

Supervised Data Augmentation



Results

Model	SectLabel Macro F1	SectLabel Micro F1	Extended Macro F1	Extended 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 + UDA_{log} †	0.784	0.906	**0.669**	**0.887**
RoBERTa-Attn Model + SDA_{log} †	**0.832**	**0.929**	0.623	0.886
SectLabel (Luong et al., 2010)‡	***0.847***	***0.934***	-	-



Scan to read the full paper!

Connect with the first author!

huangpowei@comp.nus.edu.sg