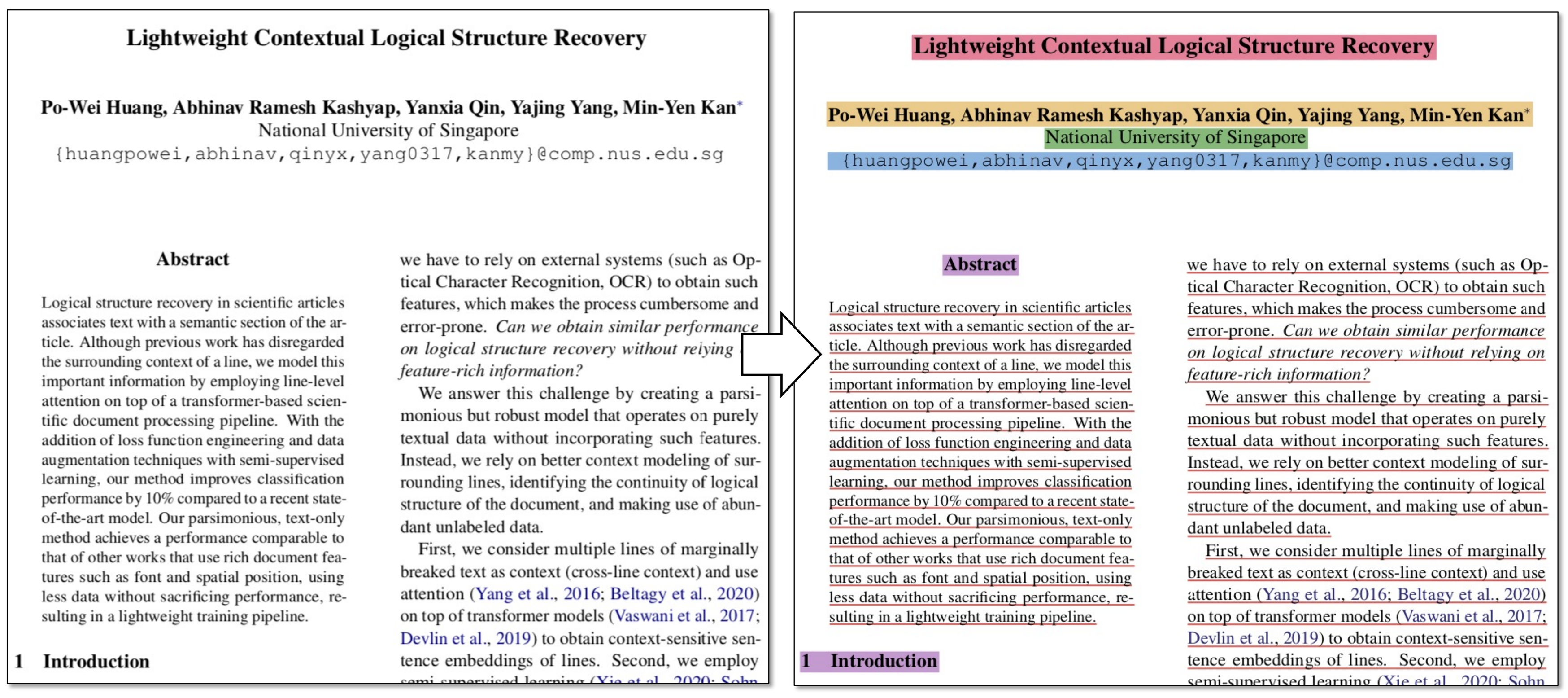
Lightweight Contextual Logical Structure Recovery

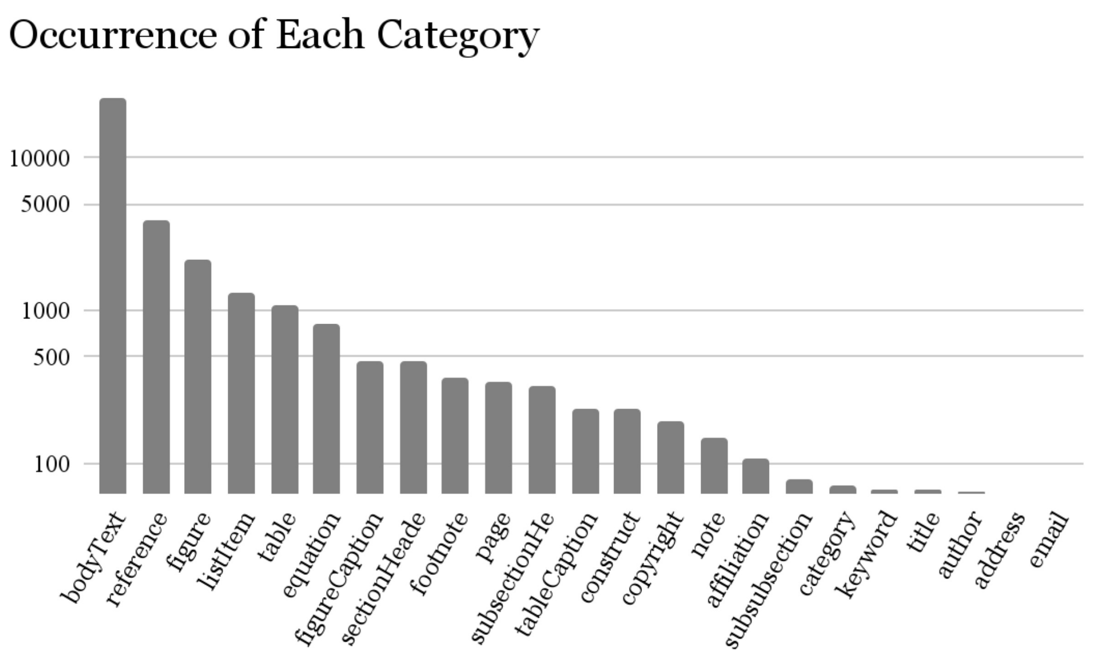
Po-Wei Huang, Abhinav Ramesh Kashyap, Yanxia Qin, Yajing Yang, Min-Yen Kan

Problem Statement



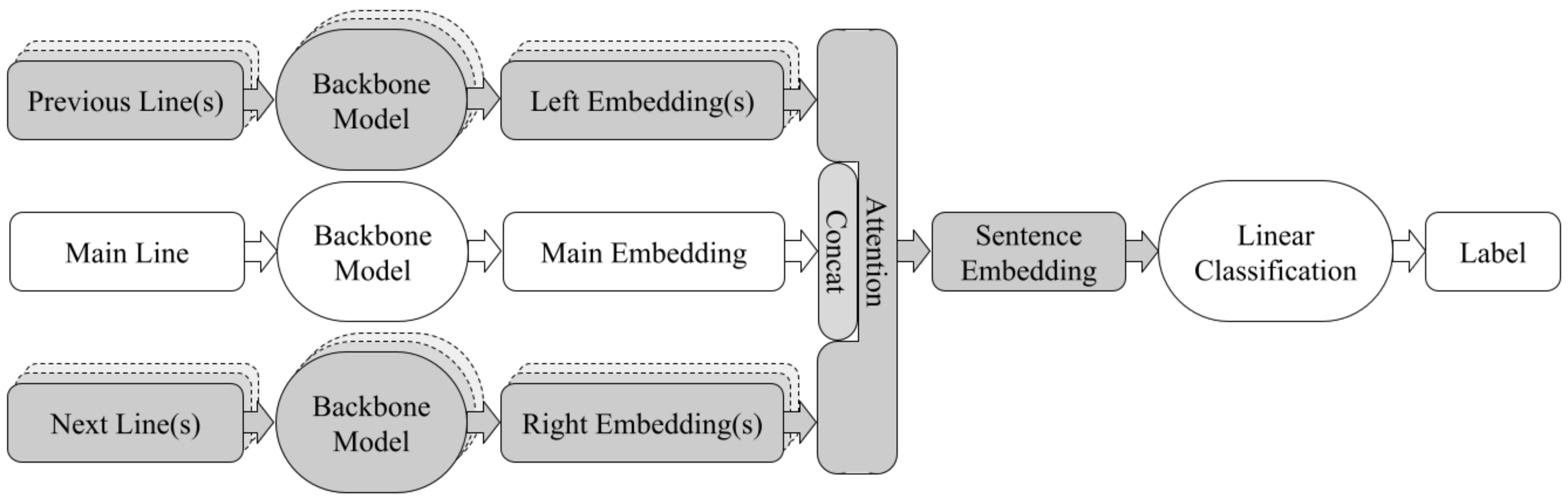
- Task: Categorize each line into 23 predefined categories that indicate the hierarchy of the document structure.
- Previous work have done this by utilizing rich text features, layout, and visional features.
- *Aim: Obtain similar performances with a contextual model on text only.*

Data



- Dataset split by document instead of by line.
- Additional labeled test dataset and unlabeled training dataset used in addition to main SectLabel dataset.

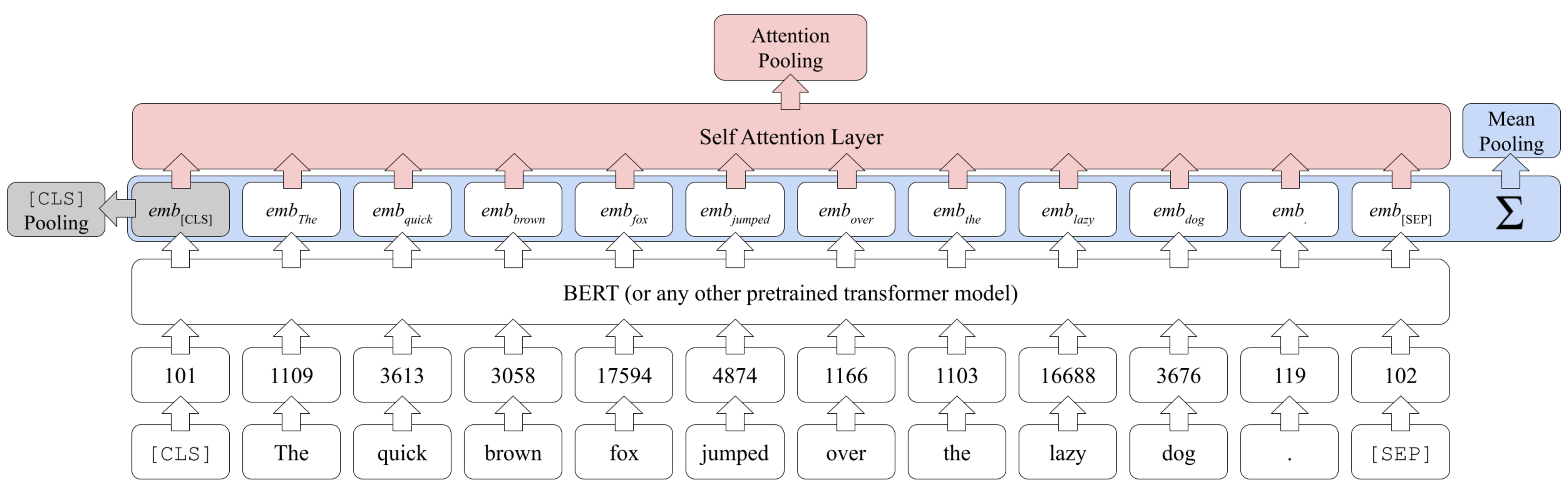
Contextual Model Construction



	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

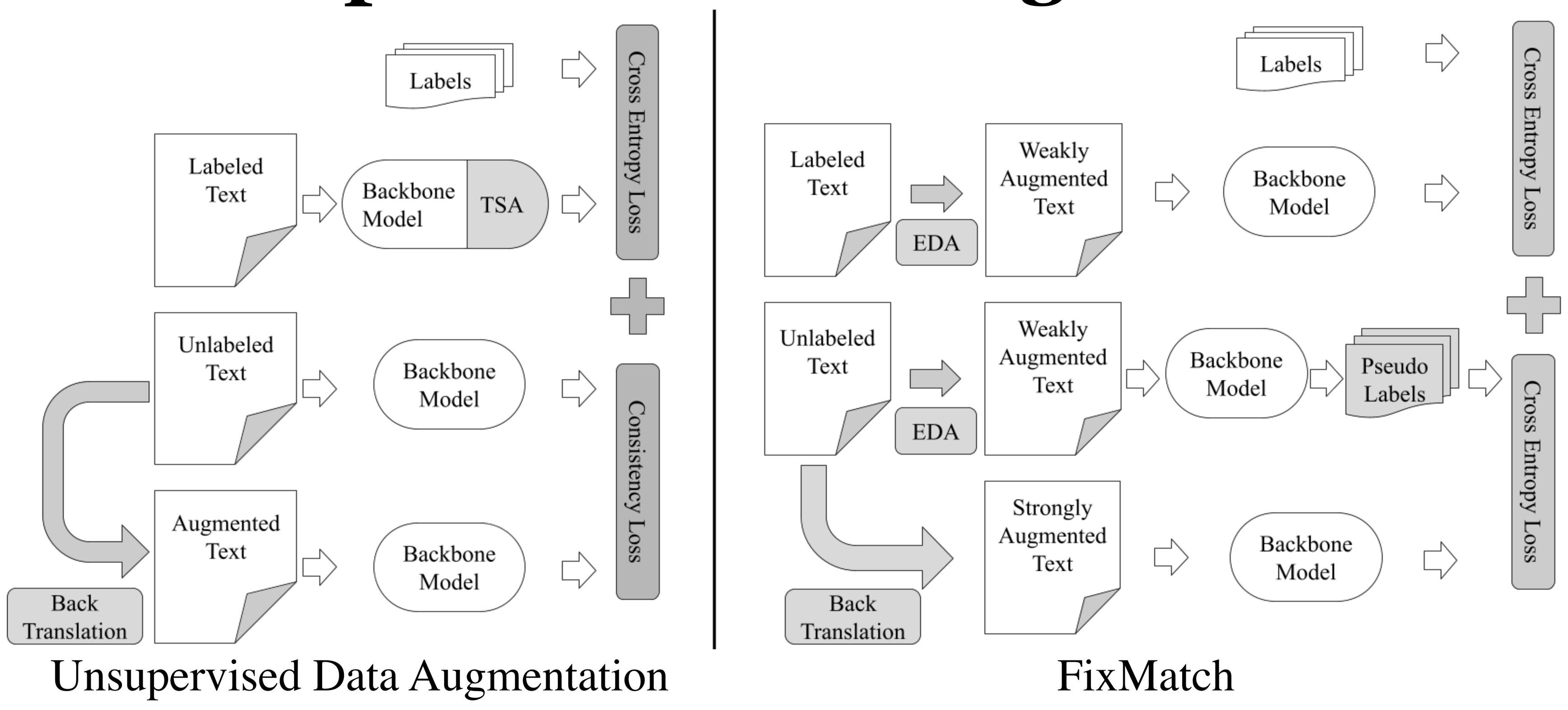
- Context of neighboring lines considered to account for the continous nature of scientific documents.
- *Sliding window attention* added as an extra layer in between sentence embedding generation and linear classification to prevent computation time increasing quadratically by document length.

Pooling for Sentence Embeddings



- Methods to generate sentence embeddings: `[CLS]` token, mean pooling, and attention pooling.

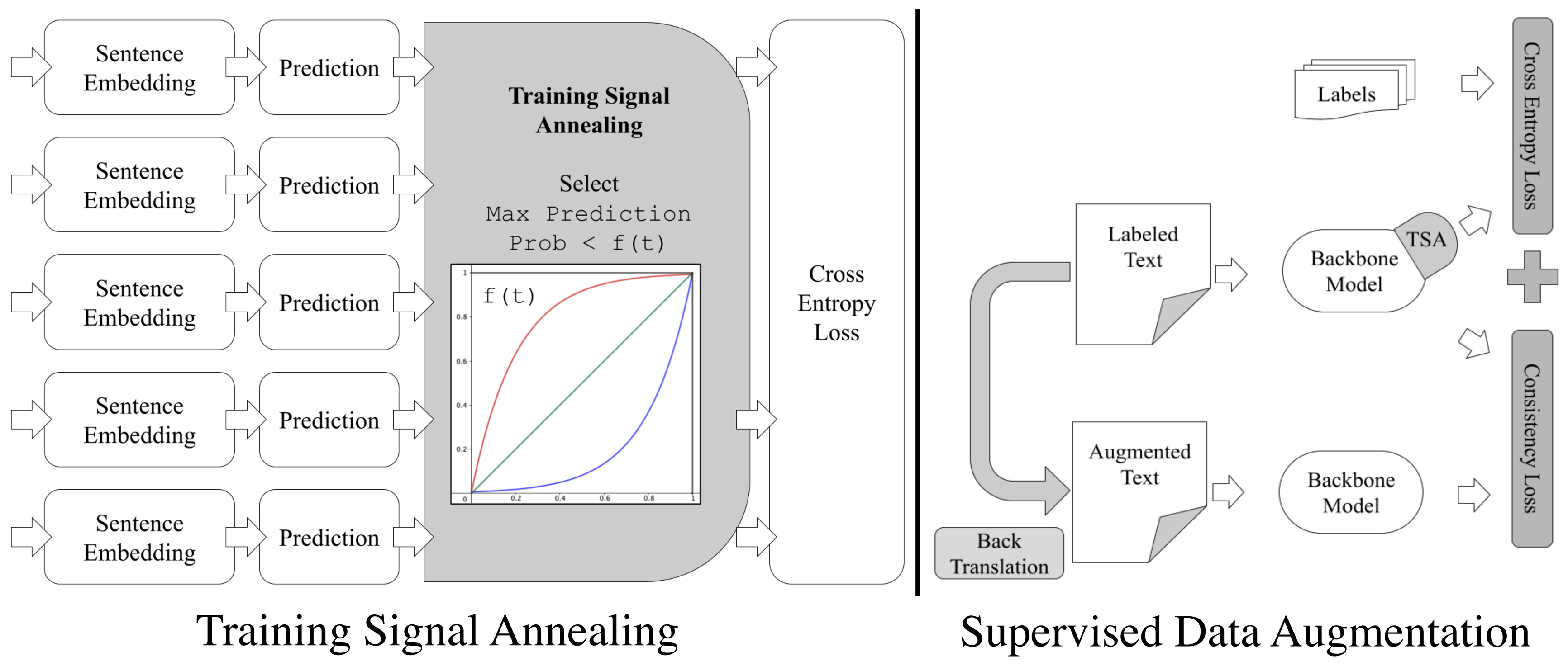
Semi-Supervised Learning



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 frameworks: Unsupervised Data Augmentation (UDA) and FixMatch.
- Data Augmentation: Back translation for strong augmentation, Easy Data Augmentation (EDA) for weak augmentation.

Loss Engineering



- Semi-supervised learning: Improves overall performance but does not improve inference on minority classes.
- Alternative: Engineer the loss term under a supervised setting to emphasizes training on minority classes.
- *Training Signal Annealing* applies a moving ceiling on the confidence of the model prediction such that only the unconfident samples are trained.
- *Supervised Data Augmentation* adds a consistency loss term to compute the divergence of the model prediction between the labelled text and its augmented version.

Results

	SectLabel		Extended	
Model	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 + 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

Scholarly Document Processing

COLING 2022