CED: Catalog Extraction from Documents

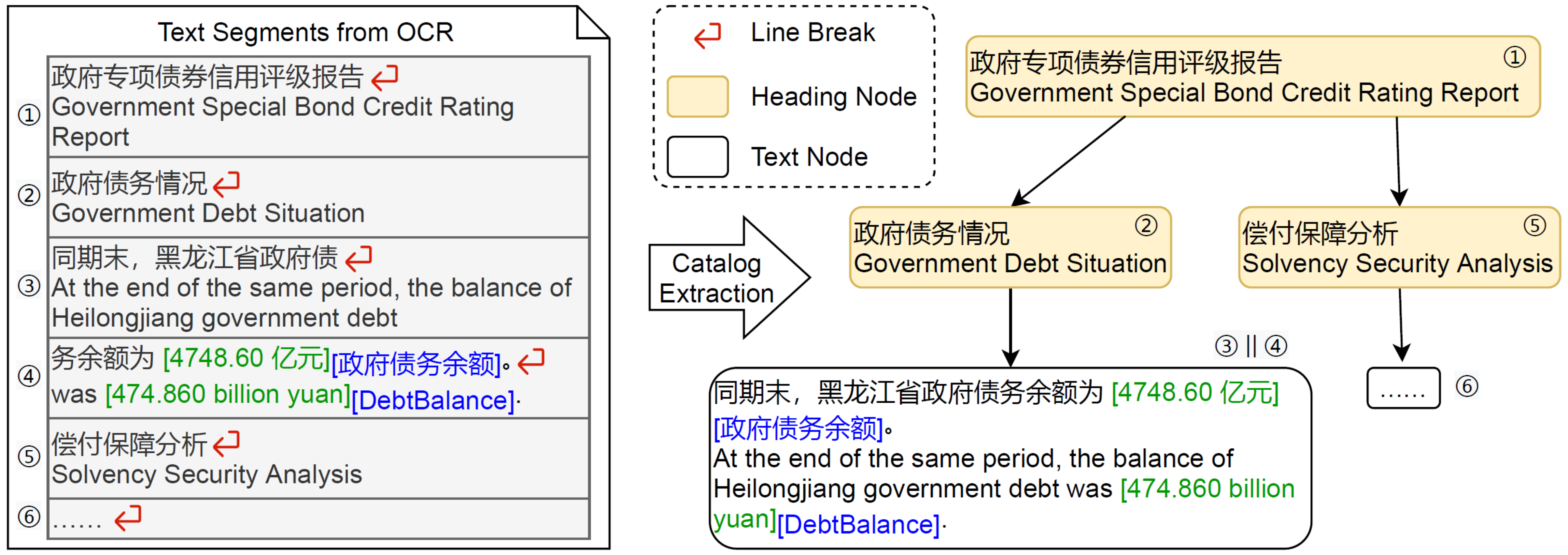
Tong Zhu[1], Guoliang Zhang[1], Zechang Li[2], Zijian Yu[1], Junfei Ren[1], Mengsong Wu[1],
Zhefeng Wang[2], Baoxing Huai[2], Pingfu Chao[1], Wenliang Chen[1] ✉
wlchen@suda.edu.cn

[1] Soochow University
[2] Huawei Cloud

Paper GitHub

Introduction

We aim to extract catalogs from documents (CED) to convert text-based documents into trees. We provide a large manually annotated dataset (ChCatExt), a transition-based model (TRACER), and a metric.



Text Segments from OCR

① 政府专项债券信用评级报告 ↩ Government Special Bond Credit Rating Report

② 政府债务情况 ↩ Government Debt Situation

③ 同期末，黑龙江省政府债 ↩ At the end of the same period, the balance of Heilongjiang government debt

④ 务余额为 [4748.60 亿元][政府债务余额]。↩ was [474.860 billion yuan][DebtBalance].

⑤ 偿付保障分析 ↩ Solvency Security Analysis

⑥ …… ↩

↩ Line Break

Heading Node

Text Node

Catalog Extraction

① 政府专项债券信用评级报告 Government Special Bond Credit Rating Report

② 政府债务情况 Government Debt Situation

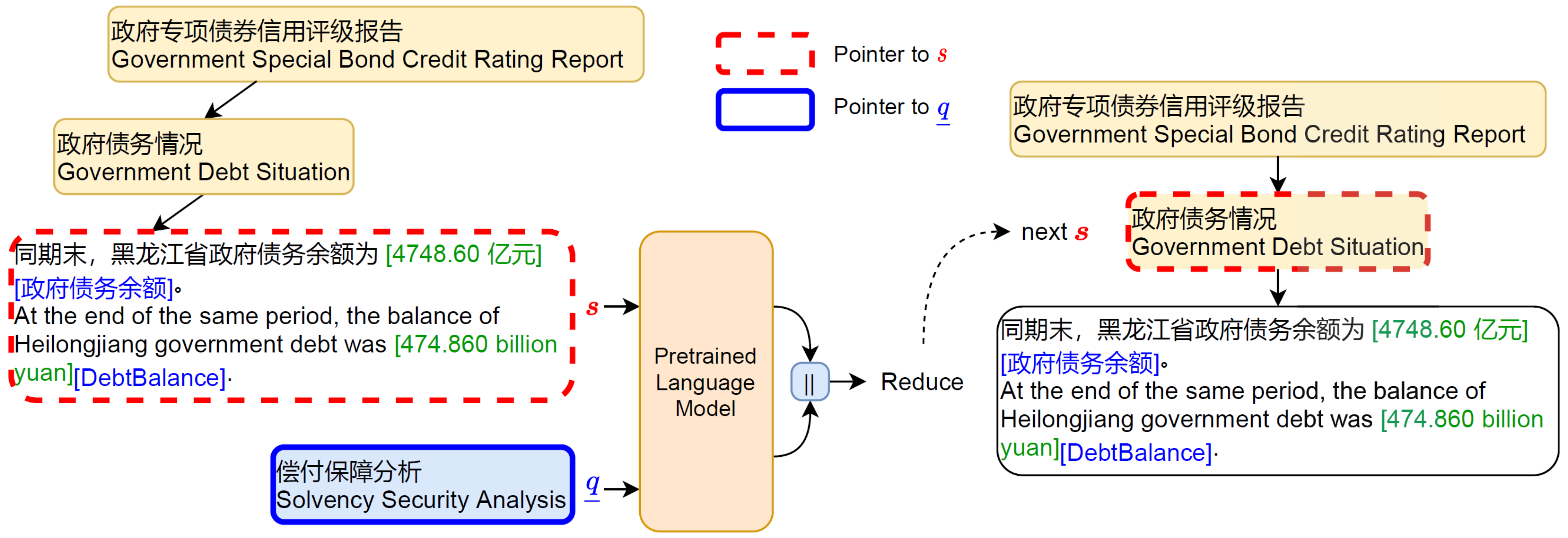
⑤ 偿付保障分析 Solvency Security Analysis

③ || ④ 同期末，黑龙江省政府债务余额为 [4748.60 亿元][政府债务余额]。At the end of the same period, the balance of Heilongjiang government debt was [474.860 billion yuan][DebtBalance].

⑥ ……

The Transition-based Method

We design a transition-based method with four actions: Sub-Heading, Sub-Text, Concat, and Reduce. Two text chunks (the stack top s and the input q) are paired to feed a BERT-family model to predict the action at each step.



政府专项债券信用评级报告 Government Special Bond Credit Rating Report

政府债务情况 Government Debt Situation

同期末，黑龙江省政府债务余额为 [4748.60 亿元][政府债务余额]。At the end of the same period, the balance of Heilongjiang government debt was [474.860 billion yuan][DebtBalance].

偿付保障分析 Solvency Security Analysis

Pointer to s

Pointer to q

Pretrained Language Model

Reduce

next s

政府专项债券信用评级报告 Government Special Bond Credit Rating Report

政府债务情况 Government Debt Situation

同期末，黑龙江省政府债务余额为 [4748.60 亿元][政府债务余额]。At the end of the same period, the balance of Heilongjiang government debt was [474.860 billion yuan][DebtBalance].

Dataset

We annotate three types of documents to construct the proposed dataset, including bid announcements, financial announcements, and credit rating reports. We also collect Wikipedia as an auto-labeled dataset for pre-training.

Source	#Docs	Avg.Length	Avg.#Nodes			Avg.Depth
			Heading	Text	Total	
BidAnn	100	1,756.76	8.04	30.61	38.65	3.00
FinAnn	300	3,504.22	12.09	52.31	64.40	3.79
CreRat	250	15,003.81	27.70	81.07	108.77	4.59
Total ChCatExt	650	7,658.30	17.47	60.03	77.50	3.98
Wiki	214,989	1,960.41	11.07	19.34	30.41	3.86

Results and Conclusion

Our model outperforms other baselines by large margins. Although Wiki pre-training is not effective in the main results, it shows good ability in domain transfer experiments.

In conclusion, we build a large dataset for automatic catalog extraction. Based on this dataset, we design a transition-based method to help address the task and obtain promising results. We hope that this task and new data could promote the development of Intelligent Document Processing.

Methods	Heading			Text			Overall		
	P	R	F1	P	R	F1	P	R	F1
Pipeline	88.637	86.595	87.601	81.627	82.475	82.047	76.837	77.338	77.085
Tagging	87.456	88.241	87.846	81.079	81.611	81.344	77.746	78.800	78.269
TRACER	**90.634**	**90.341**	**90.486**	83.031	**85.673**	**84.328**	**81.017**	**83.818**	**82.390**
w/o Constraints	89.911	89.713	89.811	82.491	84.948	83.698	80.216	83.035	81.596
TRACER w/ WikiBert	88.671	89.785	89.221	**83.308**	85.025	84.156	80.820	83.357	82.063

📦 Data, models and code are **open-sourced** at https://github.com/Spico197/CatalogExtraction