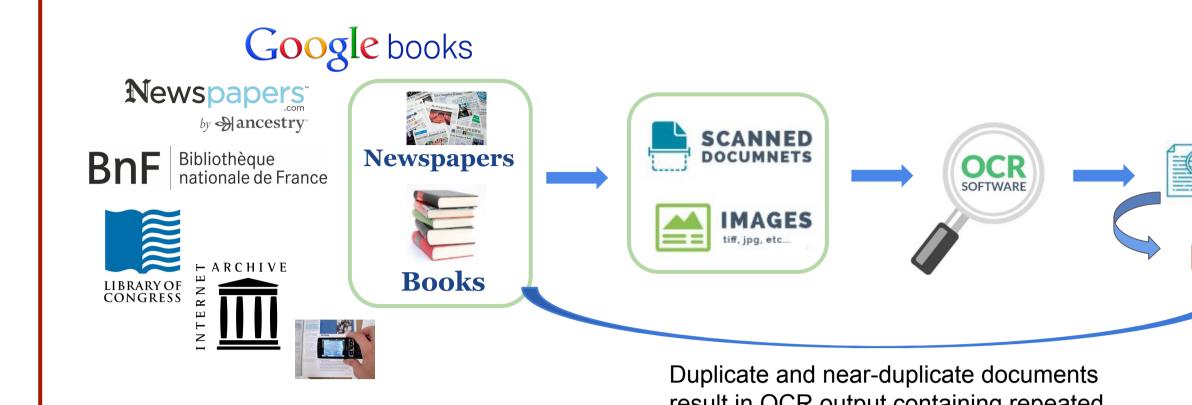


Motivation



result in OCR output containing repeated texts with various quality.

Our Goal

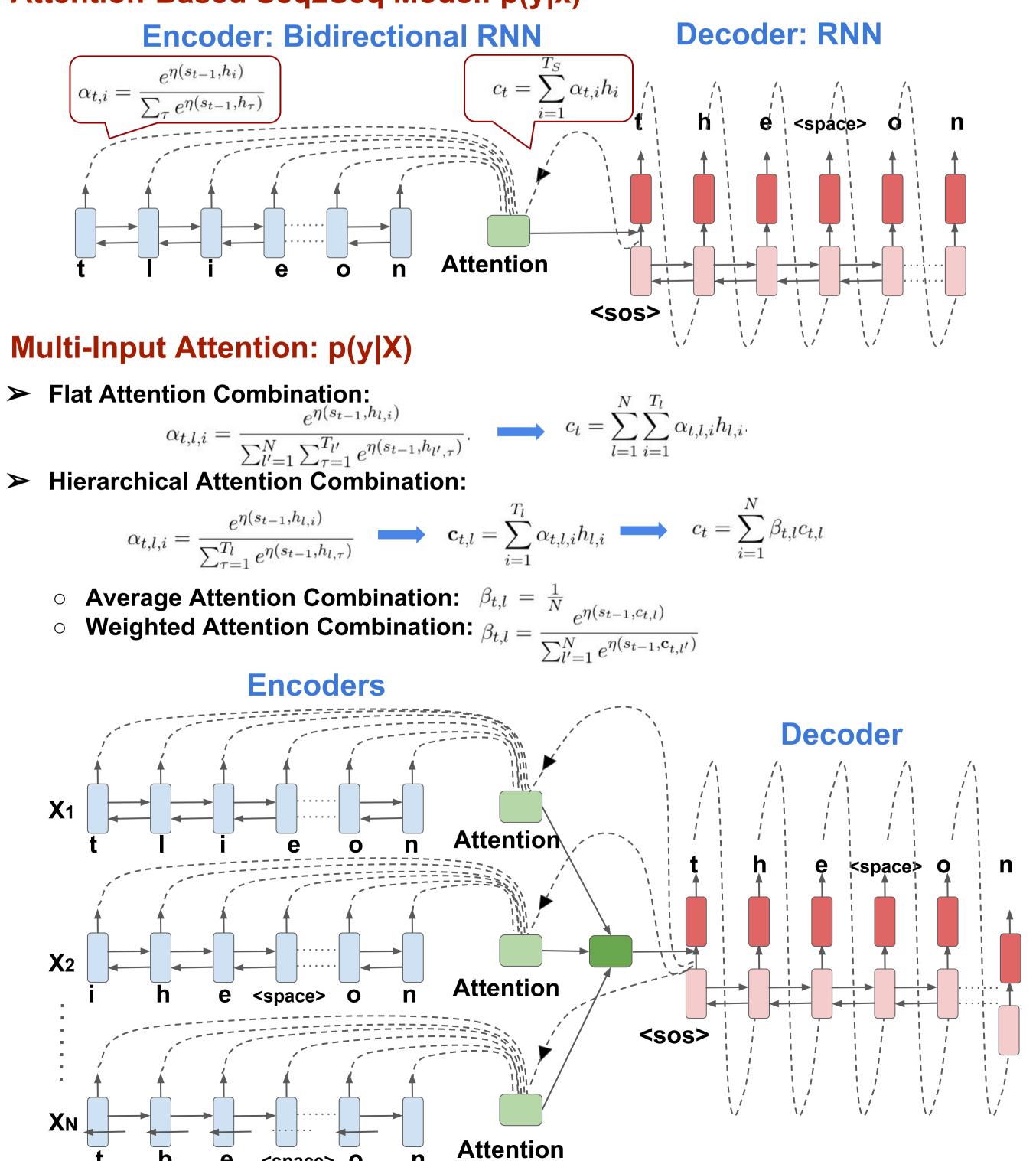
Train an *unsupervised* correction model via utilizing the duplication in OCR output that could > correct single input text sequences by mapping each erroneous OCR'd text unit to either its high-quality duplication or a consensus correction among its duplications via bootstrapping from an uniform error model. > improve the correction performance for duplicated texts by integrating multiple input sequences.

Methods

Problem Definition

Given a line of OCR'd text x, comprising the sequence of characters $[x_1, \dots, x_{T_S}]$, our goal is to map it to an error-free text $\mathbf{y} = [y_1, \dots, y_{T_T}]$ via modeling $p(\mathbf{y}|\mathbf{x})$. Given $p(\mathbf{y}|\mathbf{x})$, we also seek to model $p(\mathbf{y}|\mathbf{X})$ to search for consensus among duplicated texts \mathbf{X} , where $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N]$ are duplicated lines of OCR'd text.

Attention-Based Seq2Seq Model: p(y|x)

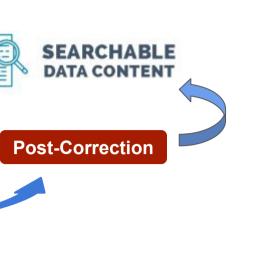


Multi-Input Attention for Unsupervised OCR Correction

College of Computer Information and Science, Northeastern University {dongrui, dasmith}@ccs.neu.edu

Datase

> Data



			➤ Multi-Input	Attentio	on Combin	ation:										
Example:						RDD	Newspa	oers		TCP Books						
to die, for I hope to go to heaven. Nor am				De	Decode		R LCE			WER	CER	LCE			'ER	
Image	would willingly me had them for such a	cause as we are fighting		1	None	0.151	49 0.04	717 0.37	'111 0 .	13799	0.1059	0.076	66 0.30	549 0.23	3495	
	for."			S	Single	0.071	99 0.03	33 0.14	906 0.	06948	0.04508	0.014	0.11	283 0.03	3392	
Manual ranscription	sorry that I have b	een slain in battle, for I			Flat 0.072				0.15818 0.062		0.05554	0.017			1079	
	eor**y that I have been slam in battle, for 1			Weighted Average		0.0688 0.042				05375 02863*	0.05516 0.04072 *	0.013 0.010			8669 092 *	
OCR output	sorry that I have b	Main Result														
	sorry tha' I have be	en s_uin in battle, f_r l		.3												
stics of Dat	asets:		Decode	Model		RDD Newspapers TCP Books							Books			
Dataset	# Lines with	# Lines w/manual		Decoue	wode	;I	CER	LCER	WER	LW	/ER C	ER	LCER	WER	LWER	
	w/manual	& witness			None		0.18133	0.13552	0.4178	0.3	1544 0.	1067	0.088	0.31734	0.27227	
RDD	2.2M	0.95M (43%)			Seq2Seq-Supe		0.09044	0.04469	0.1781	2 0.09	9063 0.0)4944	0.01498	0.12186	0.035	
TCP	8.6M	5.5M (64%)			Seq2Seq-N		0.10524	0.05565	0.206			08704	0.05889	0.25994	0.15725	
nony Poculto					Seq2Seq-	-	0.16136	0.11986	0.3580)9551	0.0616	0.27845	0.18221	
lary ites	nary Results					_	0.11037	0.06149	0.2275			07196	0.03684	0.21711	0.11233	
e Input Cor	e Input Correction Model:				Seq2Seq-Boots											
•put • • •							0.15507	0.13552	0.3465			0862	0.088	0.33983	0.27227	
Mode	el CER	WER			Majority \		0.16285	0.13552	0.4006			1096	0.088	0.34151	0.27227	
None	0.18133	0.41780		Multi	Seq2Seq-S	Super	0.07731	0.03634	0.1539	3 0.07	7269 0.0	04668	0.01252	0.11236	0.02667	
PCRF(order=	=5, w=4) 0.11403	0.25116		IVIGIU	Seq2Seq-N	Noisy	0.09203*	0.04554*	0.1794	0.09	9269 0.0	08317	0.05588	0.24824	0.14885	
PCRF(order=					Seq2Seq-	-Syn	0.12948	0.09112	0.2890	1 0.19	9977 0.0	08506	0.05002	0.24942	0.15169	
Attn-Seq2	2Seq 0.11028 [*]	0.23405*			Seq2Seq-E	Boots	0.09435	0.04976	0.1968	1 0.10	0.0 0.0	6824*	0.03343*	0.20325*	0.09995	

> Stati

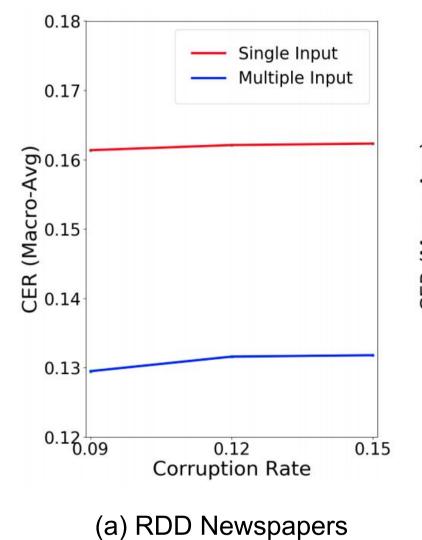
t			> Multi-Input	Attentio	n Combin	ation										
Example:							RDD	Newsp	apers		TCP Books					
_	to die, for I hope to g sorry that I have been	o die, for I hope to go to heaven. Nor am I forry that I have been flain in battle, for I			ecode	CE			VER	LWER	CE	R LO	CER	WER	LWE	R
Image		rifice a decan lives if I		1	None	0.151	49 0.047	'17 0.	37111	0.13799	0.10	0.0	7666	0.30549	0.2349	95
	for."			Single		0.071	99 0.03	3 0.	14906	0.06948	0.04	508 0.0	1407	0.11283	0.0339	92
Manual ranscription	sorry that I have been slain in battle, for I			Flat		0.072 0.068			15818 15221	0.06241* 0.05375	0.05		1727 1392*	0.13487	0.0407 0.0366	
	eor**y that I have be	eor**y that I have been slam in battle, for 1			eighted /erage	0.000			09397	0.03373	0.03			0.133 0.09786 *	0.0300	
OCR output	-	en slain in battle, for I	Main Results		lage								<u> </u>			
	sorry tha' I have be	en s _u in in battle, f_r l														
istics of Datasets:					Madal		RDD Newspapers			-			TCP Books			
Dataset	# Lines with w/manual	# Lines w/manual & witness)ecode	Mode		CER				WER	CER	LCE		ER	LWE
RDD	2.2M	0.95M (43%)			None		0.18133	0.13552	2 0.4	178 0.	31544	0.1067	30.0	38 0.3	1734	0.2722
ТСР	8.6M	5.5M (64%)			Seq2Seq-S	Super	0.09044	0.04469	9 0.17	7812 0.	09063	0.04944	0.014	198 0.1 2	2186	0.035
<u> </u>			Image: Market interview Imarkt Image: Market interview	206 0.	11416	0.08704	0.058	889 0.2	5994	0.1572						
nary Res	ults			Single	Seq2Seq-	-Syn	0.16136	0.11986	6 0.35	5802 0.	26547	0.09551	0.06	16 0.2	7845	0.1822
					Seq2Seq-	Boots	0.11037	0.06149	9 0.2	275 0.	13123	0.07196	0.036	684 0.2	1711	0.1123
e Input Coi	rrection Model:				LMR		0.15507	0.13552	2 0.34	4653 0.	31544	0.10862	30.0	38 0.3	3983	0.2722
Mada					Majority \	/ote	0.16285	0.13552	2 0.40	0063 0.	31544	0.11096	0.08	38 0.3 [,]	1151	0.2722
Mode		WER			Seq2Seq-S		0.07731	0.03634			07269	0.04668	0.012			0.026
None		0.41780		Multi	Seq2Seq-I		0.09203*	0.04554			09269	0.08317	0.055			0.1488
PCRF(order= PCRF(order=		0.25116														
Attn-Seq2					Seq2Seq-		0.12948	0.09112			19977	0.08506	0.050			0.151
					Seq2Seq-E	Boots	0.09435	0.04976	6 0.19	9681 0.	10604	0.06824*	0.033	43* 0.20	325*	0.0999

Prelimi

➤ Single

t			> Multi-Input	Attentic	on Combir	nation	:										
a Example:				RDD Newspapers							TCP Books						
to die, for I hope to go to heaven. Nor am I			D	Decode		R LCE		WER	LWER	CE	R LC	ER W	ER LV	VER			
Image	would willingly me	rifice a dozen lives if I			None	0.151	49 0.047	717	0.37111	0.13799	0.10	0.07	666 0.3	0549 0.2	23495		
	for."			S	Single	0.071	99 0.03	33	0.14906	0.06948	0.04	508 0.01	407 0.1	1283 0.0)3392		
Manual Transcription	sorry that I have be			Flat	0.072			0.15818 0.15221	0.06241* 0.05375	0.05				04079			
	eor**y that I have been slam in battle, for 1			Weighted		0.068 0.042			0.1 5221 0.09397	0.05375	0.05						
OCR output	sorry that I have be	en slain in battle, for I	Main Result		verage	0.042	10 0.013	999	0.09397	0.02003	0.040	J72 0.01	021 0.0	9700 0.0	2092*		
sorry tha' I have been s_uin in battle, f_r I tistics of Datasets:					_	RDD Newspapers						Books					
Dataset	# Lines with w/manual	# Lines w/manual & witness		Decode	Mode)	CER	LCE			VER	CER	LCER	WER	LV		
RDD	2.2M	0.95M (43%)			None		0.18133	0.135			1544	0.1067	0.088	0.31734	_		
ТСР	8.6M	5.5M (64%)			Seq2Seq-S	Super	0.09044	0.044	169 0.1	7812 0.0	9063	0.04944	0.01498	0.12186	0.		
inary Results				Single	Seq2Seq-l Seq2Seq·		0.10524 0.16136	0.055 0.119			1416 6547	0.08704 0.09551	0.05889 0.0616	0.25994 0.27845	0.1		
-					Seq2Seq-Boots		0.11037	0.11037 0.06149		49 0.2275 0.13		0.07196	0.03684 0.2		0.1		
e Input Co	rrection Model:				LMR		0.15507	0.135	552 0.3	4653 0.3	1544	0.10862	0.088	0.33983	0.2		
Mode	el CER	WER			Majority V	Vote	0.16285	0.135	552 0.4	0063 0.3	1544	0.11096	0.088	0.34151	0.2		
None		0.41780		N AI.t.'	Seq2Seq-S	Super	0.07731	0.036	634 0.1	5393 0.0	7269	0.04668	0.01252	0.11236	0.0		
PCRF(order=		0.25116		Multi	Seq2Seq-l	Noisy	0.09203*	0.045	54* 0.1	1794 0.0	9269	0.08317	0.05588	0.24824	0.1		
PCRF(order=		0.25617			Seq2Seq	•	0.12948	0.091	12 0.2	8901 0.1	9977	0.08506	0.05002	0.24942	0.1		
Attn-Seq2	2Seq 0.11028*	0.23405*			Seq2Seq-I		0.09435	0.049			0604	0.06824*	0.03343*	0.20325*			

> Does Corruption Rate Affect Synthetic Training?



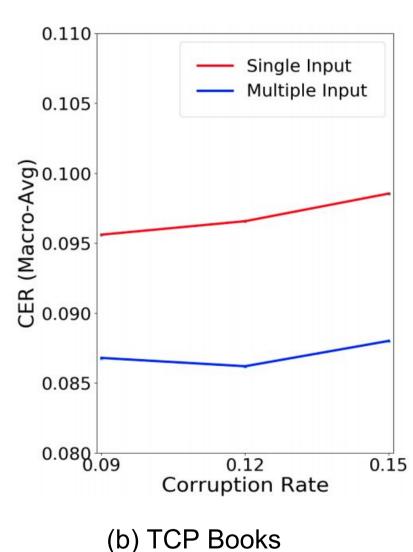


Figure 2: Performance of Seq2Seq-Syn trained on synthetic data with different corrpution rates.

Training

- Supervised Training (Seq2Seq-Super): map each OCR'd line into the corresponding manual transcription.
- > Unsupervised Training:
 - Noisy Training (Seq2Seq-Noisy)
 - Rank the duplicated texts by scores assigned by a language model. Train a correction model to map the OCR'd line to its high-quality duplication.
 - Synthetic Training (Seq2Seq-Syn)
 - Train a correction model to recover a manually corrupted corpus. Synthetic Training with Boostrap (Seq2Seq-Boots)
 - Utilize the multi-input attention mechanism to generate a high-quality consensus correction for each OCR'd line with duplicated texts via the model with synthetic training. Train a correction model to transform the OCR'd lines to their consensus corrections.

Rui Dong, David Smith

Results

Further Experiments

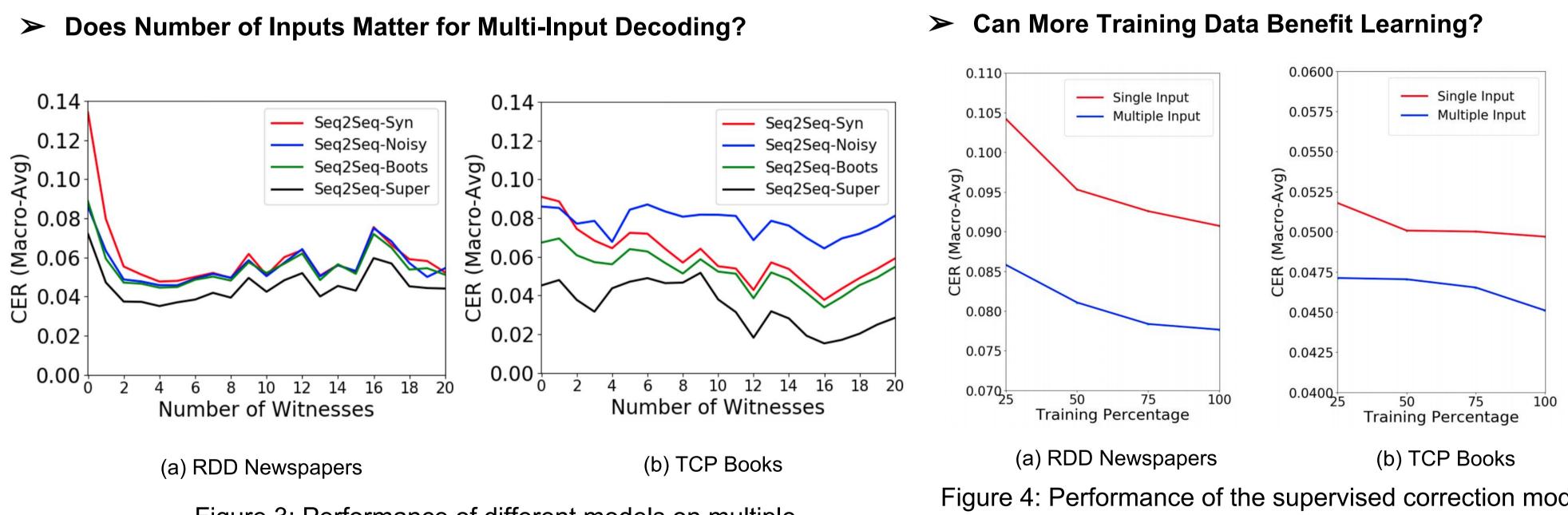


Figure 3: Performance of different models on multiple decoding of lines with different number of witnesses.

Our Contributions:

- \succ a scalable framework needing **no supervision** from human annotations to train the correction model
- a multi-input attention mechanism incorporating aligning, correcting,
- a method that corrects text either with or without duplicated versions,

Figure 4: Performance of the supervised correction model trained on different proportions of the RDD newspapers and TCP books Dataset.

Conclusions

and voting on multiple sequences simultaneously for consensus decoding, which is more efficient and effective than existing ensemble methods while most existing methods can only deal with one of these cases.