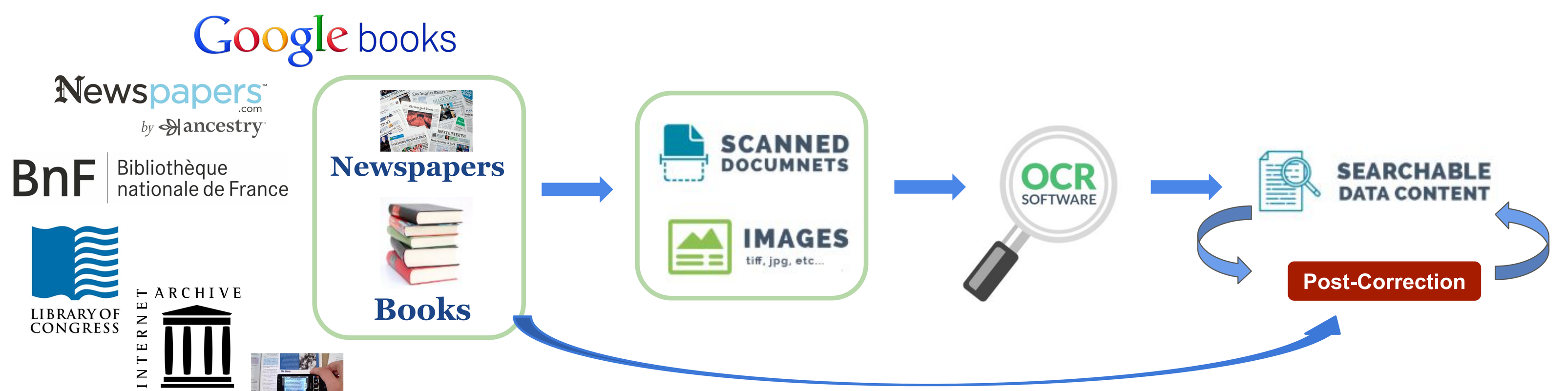
Multi-Input Attention for Unsupervised OCR Correction

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

Motivation



Duplicate and near-duplicate documents 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

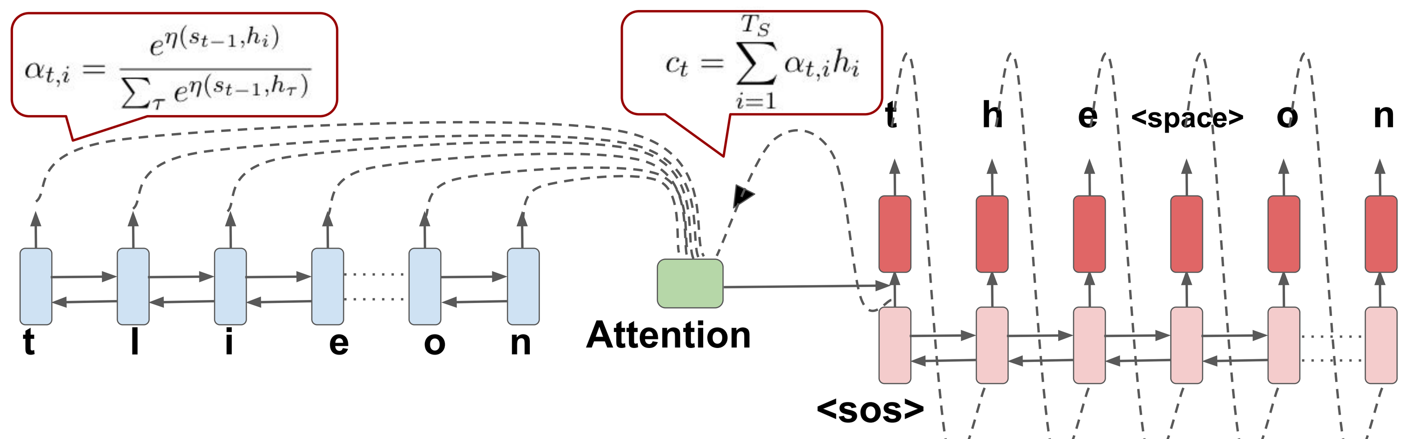
- ➢ correct single input text sequences by mapping each erroneous OCR'd text unit to either its high-quality duplication or a consensus correction among its duplications via bootstrapping from an uniform error model.
- ➢ improve the correction performance for duplicated texts by integrating multiple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

Methods

Problem Definition

Given a line of OCR'd text $\mathbf{x}$, comprising the sequence of characters $[x_1, \cdots, x_{T_S}]$, our goal is to map it to an error-free text $\mathbf{y} = [y_1, \cdots, y_{T_T}]$ 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: p(y|X)

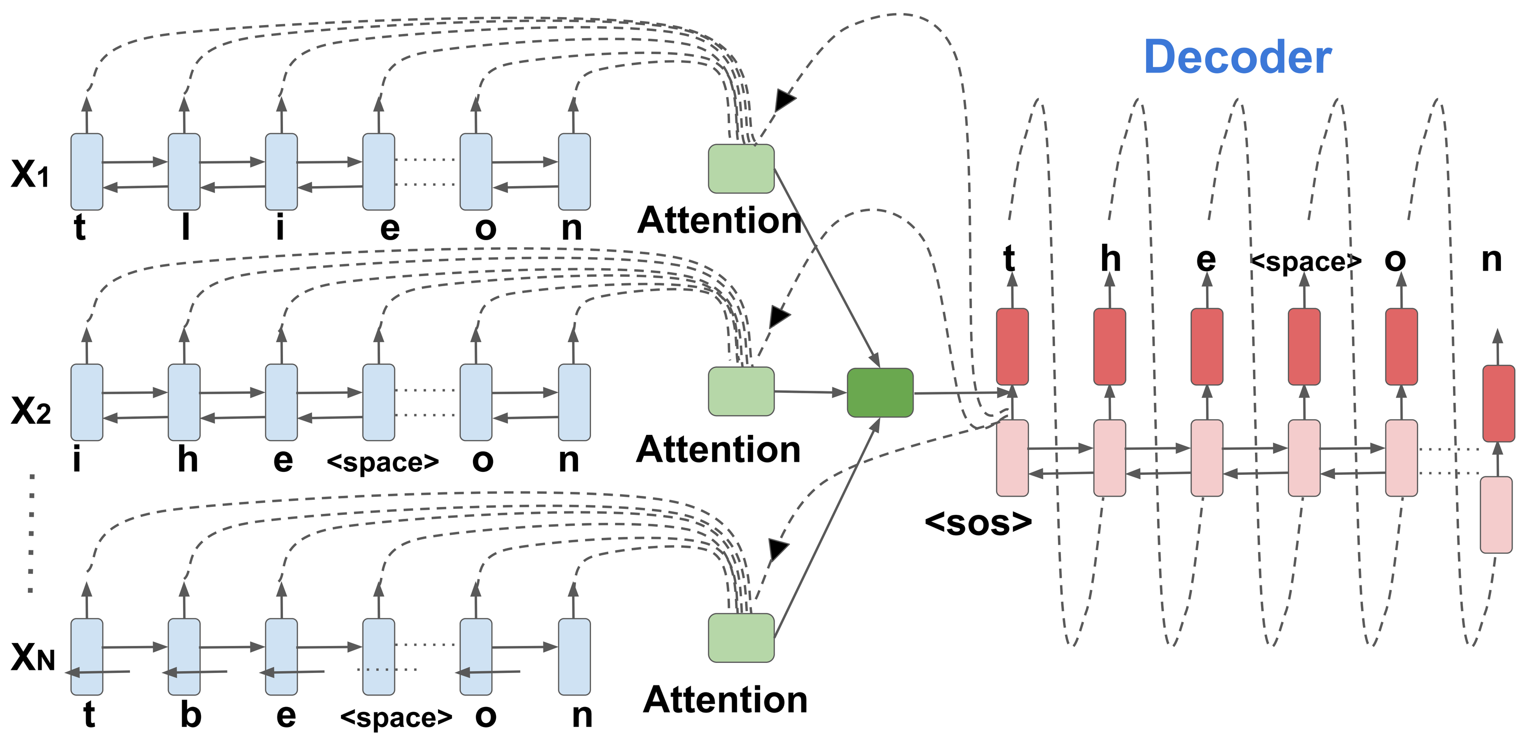
➢ **Flat Attention Combination:**

$$\alpha_{t,l,i} = \frac{e^{\eta(s_{t-1},h_{l,i})}}{\sum_{l'=1}^{N}\sum_{\tau=1}^{T_{l'}} e^{\eta(s_{t-1},h_{l',\tau})}}. \rightarrow c_t = \sum_{l=1}^{N}\sum_{i=1}^{T_l} \alpha_{t,l,i} h_{l,i}.$$

➢ **Hierarchical Attention Combination:**

$$\alpha_{t,l,i} = \frac{e^{\eta(s_{t-1},h_{l,i})}}{\sum_{\tau=1}^{T_l} e^{\eta(s_{t-1},h_{l,\tau})}} \rightarrow \mathbf{c}_{t,l} = \sum_{i=1}^{T_l} \alpha_{t,l,i} h_{l,i} \rightarrow c_t = \sum_{i=1}^{N} \beta_{t,l} c_{t,l}$$

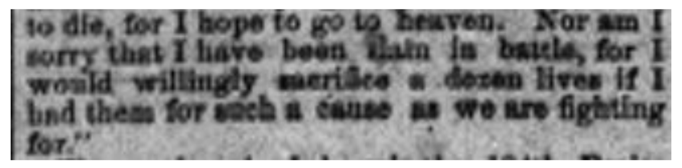
- ○ **Average Attention Combination:** $\beta_{t,l} = \frac{1}{N}$
- ○ **Weighted Attention Combination:** $\beta_{t,l} = \frac{e^{\eta(s_{t-1},c_{t,l})}}{\sum_{l'=1}^{N} e^{\eta(s_{t-1},\mathbf{c}_{t,l'})}}$



Results

Dataset

➢ **Data Example:**

Image	
Manual Transcription	sorry that I have been slain in battle, for I
OCR output	eor**y that I have been slam in battle, for 1
	sorry that I have been slain in battle, for I
	sorry tha' I have been s_uin in battle, f_r I

➢ **Statistics of Datasets:**

Dataset	# Lines with w/manual	# Lines w/manual & witness
RDD	2.2M	0.95M (43%)
TCP	8.6M	5.5M (64%)

Preliminary Results

➢ **Single Input Correction Model:**

Model	CER	WER
None	0.18133	0.41780
PCRF(order=5, w=4)	0.11403	0.25116
PCRF(order=5, w=6)	0.11535	0.25617
Attn-Seq2Seq	**0.11028***	**0.23405***

➢ **Multi-Input Attention Combination:**

Decode	RDD Newspapers				TCP Books			
	CER	LCER	WER	LWER	CER	LCER	WER	LWER
None	0.15149	0.04717	0.37111	0.13799	0.1059	0.07666	0.30549	0.23495
Single	0.07199	0.033	0.14906	0.06948	0.04508	0.01407	0.11283	0.03392
Flat	0.07238	0.02904*	0.15818	0.06241*	0.05554	0.01727	0.13487	0.04079
Weighted	0.06882*	0.02145*	0.15221	0.05375	0.05516	0.01392*	0.133	0.03669
Average	**0.04210***	**0.01399***	**0.09397**	**0.02863***	**0.04072***	**0.01021***	**0.09786***	**0.02092***

Main Results

Decode	Model	RDD Newspapers				TCP Books			
		CER	LCER	WER	LWER	CER	LCER	WER	LWER
	None	0.18133	0.13552	0.4178	0.31544	0.1067	0.088	0.31734	0.27227
Single	Seq2Seq-Super	0.09044	0.04469	0.17812	0.09063	0.04944	0.01498	0.12186	0.035
	Seq2Seq-Noisy	0.10524	0.05565	0.206	0.11416	0.08704	0.05889	0.25994	0.15725
	Seq2Seq-Syn	0.16136	0.11986	0.35802	0.26547	0.09551	0.0616	0.27845	0.18221
	Seq2Seq-Boots	0.11037	0.06149	0.2275	0.13123	0.07196	0.03684	0.21711	0.11233
Multi	LMR	0.15507	0.13552	0.34653	0.31544	0.10862	0.088	0.33983	0.27227
	Majority Vote	0.16285	0.13552	0.40063	0.31544	0.11096	0.088	0.34151	0.27227
	Seq2Seq-Super	0.07731	0.03634	0.15393	0.07269	0.04668	0.01252	0.11236	0.02667
	Seq2Seq-Noisy	**0.09203***	**0.04554***	**0.1794**	**0.09269**	0.08317	0.05588	0.24824	0.14885
	Seq2Seq-Syn	0.12948	0.09112	0.28901	0.19977	0.08506	0.05002	0.24942	0.15169
	Seq2Seq-Boots	0.09435	0.04976	0.19681	0.10604	**0.06824***	**0.03343***	**0.20325***	**0.09995***

Further Experiments

➢ **Does Corruption Rate Affect Synthetic Training?**

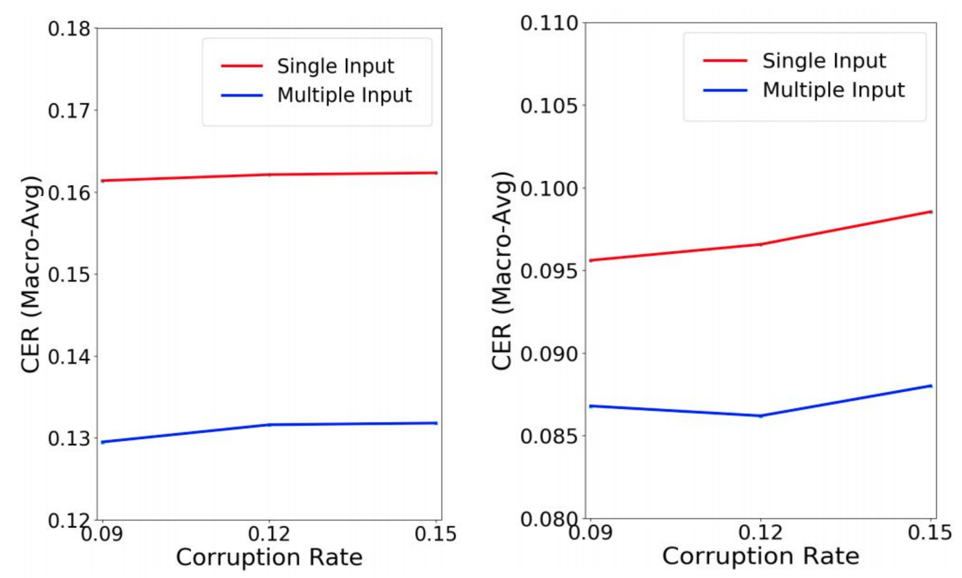


Figure 2: Performance of Seq2Seq-Syn trained on synthetic data with different corrpution rates.

➢ **Does Number of Inputs Matter for Multi-Input Decoding?**

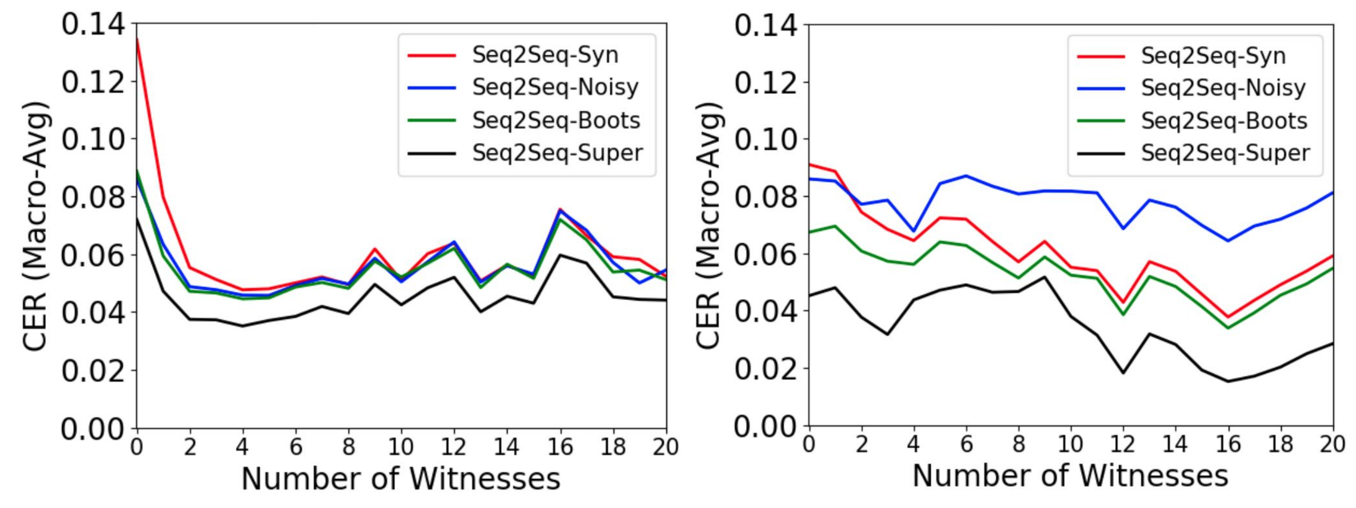


Figure 3: Performance of different models on multiple decoding of lines with different number of witnesses.

➢ **Can More Training Data Benefit Learning?**

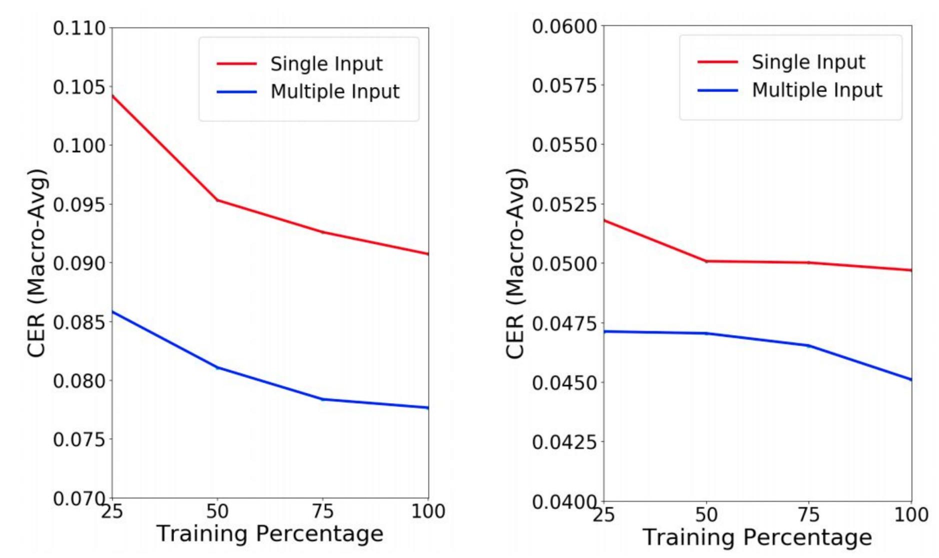


Figure 4: Performance of the supervised correction model trained on different proportions of the RDD newspapers and TCP books Dataset.

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.

Conclusions

Our Contributions:

- ➢ a scalable framework needing **no supervision** from human annotations to train the correction model
- ➢ a **multi-input attention mechanism** incorporating ***aligning, correcting, and voting*** on multiple sequences simultaneously for consensus decoding, which is more efficient and effective than existing ensemble methods
- ➢ a method that corrects text either **with or without duplicated versions**, while most existing methods can only deal with one of these cases.