HIBERT: Document Level Pre-training of Hierarchical Bidirectional Transformers for Document Summarization

Xingxing Zhang, Furu Wei and Ming Zhou

Microsoft Research Asia, Beijing, China

31 July, 2019

Microsoft Research
微软亚洲研究院

Document Summarization

Share 225 shares 242 View comments

Here was a glimpse into the exciting future Ole Gunnar Solskjaer has mapped out for his young Manchester United side.

It's only mid-July and this was an outnumbered Leeds side, but there were encouraging signs all the same.

A first senior goal for Mason Greenwood, a wonderful effort from Marcus Rashford and a cute piece of skill from Tahith Chong to earn a penalty for United's fourth goal were the standout moments.

Document Summarization

- Manchester United won the bragging rights over Leeds with a 4-0 win in Perth
- Mason Greenwood got them off to a perfect start to score inside 10 minutes
- Marcus Rashford doubled the lead just before the half-hour mark in Perth
- Phil Jones made it 3-0 five minutes after coming on as a substitute at half-time
- Tahith Chong superbly won a penalty which Anthony Martial easily converted

Related Work: Summarization

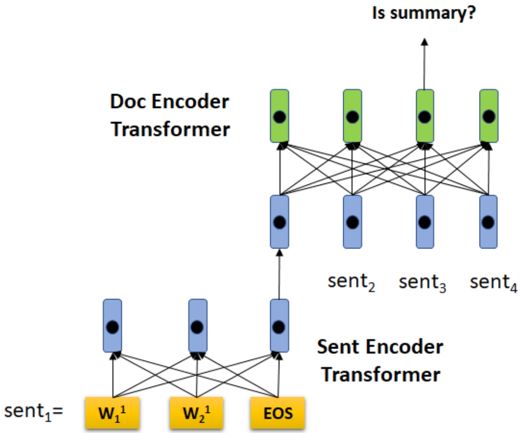
Extractive Summarization (This work)

- Sentence Ranking/Classification
- Sparse Features: Nenkova and McKeown (2011)
- Hierarchical CNN/LSTM:
    - Cheng and Lapata, (2016)
    - Narayan et al., (2018); Dong et al., (2018)
    - Zhang et al., (2018); Zhou et al., (2018)

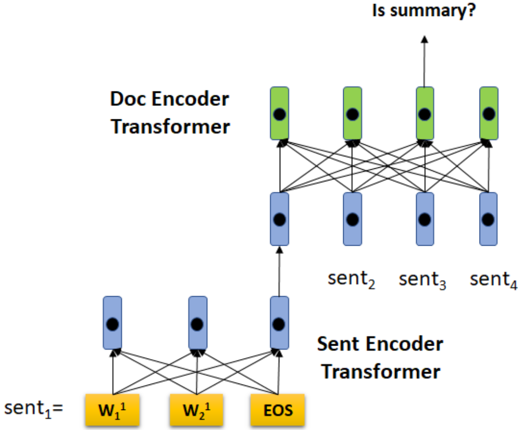
Abstractive Summarization

- Seq2Seq: Copy-Generator (See et al., 2017)
- Reinforce (Paulus et al., 2017)
- Extract-Generate
    - Chen and Bansal, (2018); Gehrmann et al., (2018)

Extractive Summarization with Hierarchical Transformers

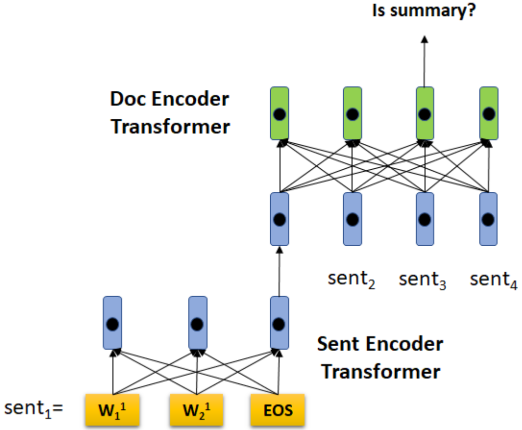


Extractive Summarization with Hierarchical Transformers



- Why not train with extractive labels?

Extractive Summarization with Hierarchical Transformers

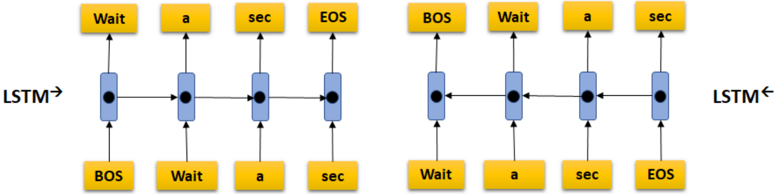


- Why not train with extractive labels?
- Pre-training Hierarchical Transformers (i.e. Document Encoders) may help. How?

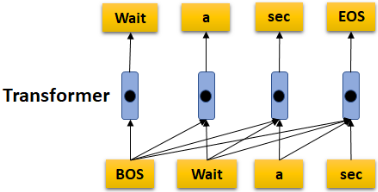
Pre-training of Nonhierarchical/Sentence Encoders

Language Modeling as Training Objective

- ELMo (Peters et al., 2018)



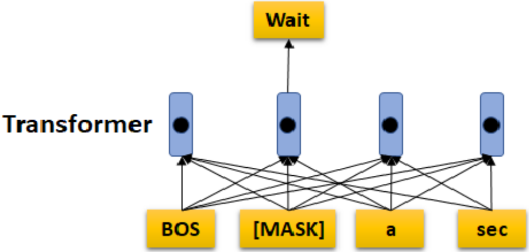
- GPT (Radford et al., 2018)



Pre-training of Nonhierarchical/Sentence Encoders

Masked Language Modeling (*Cloze*) as Training Objective

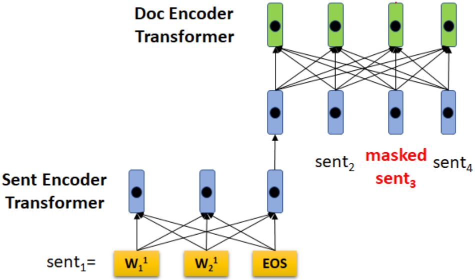
- *Cloze* (Taylor, 1953)
- BERT (Devlin et al., 2019)



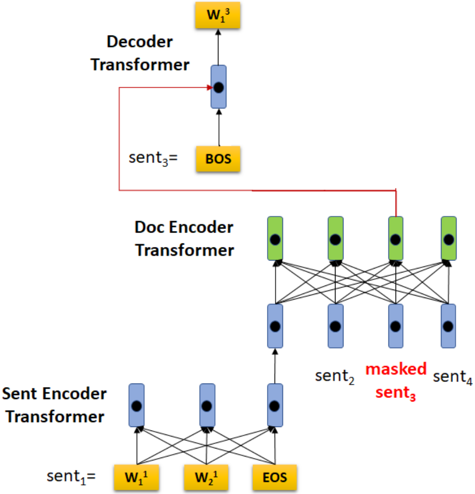
- Obtained better results than L2R or R2L models

Pre-training of HIBERT

HIerachical **B**idirectional **E**ncoder **R**epresentations from **T**ransformers

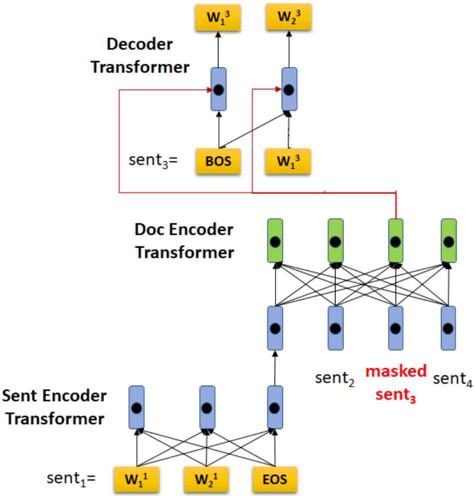


HIerachical **B**idirectional **E**ncoder **R**epresentations from **T**ransformers



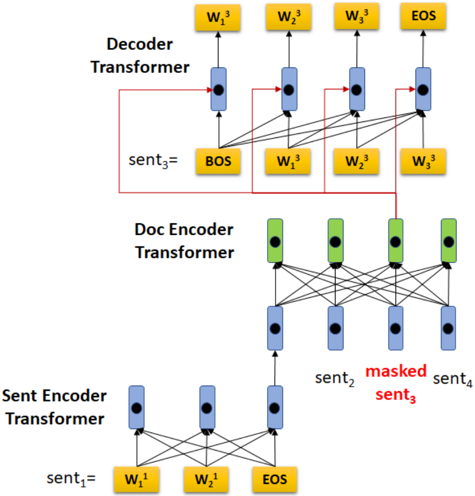
Pre-training of HIBERT

HIerachical **B**idirectional **E**ncoder **R**epresentations from **T**ransformers



Pre-training of HIBERT

HIerachical **B**idirectional **E**ncoder **R**epresentations from **T**ransformers



Pre-training of HIBERT

Document Masking

`William Shakespeare is a poet . He died in 1616 . He is regarded as the greatest writer .`

- Randomly select 15% of the sentences in a doc

Pre-training of HIBERT

Document Masking

`William Shakespeare is a poet . He died in 1616 . He is regarded as the greatest writer .`

- Randomly select 15% of the sentences in a doc
- **MASK:** 80% of cases, we mask them

  `William Shakespeare is a poet . [MASK] [MASK] [MASK] [MASK] [MASK] He is regarded as the greatest writer .`

Document Masking

`William Shakespeare is a poet . He died in 1616 . He is regarded as the greatest writer .`

- Randomly select 15% of the sentences in a doc
- **MASK:** 80% of cases, we mask them

  `William Shakespeare is a poet . [MASK] [MASK] [MASK] [MASK] [MASK] He is regarded as the greatest writer .`
- **KEEP:** 10% of cases, we keep them unchanged

Pre-training of HIBERT

Document Masking

`William Shakespeare is a poet . He died in 1616 . He is regarded as the greatest writer .`

- Randomly select 15% of the sentences in a doc
- **MASK:** 80% of cases, we mask them

  `William Shakespeare is a poet . [MASK] [MASK] [MASK] [MASK] [MASK] He is regarded as the greatest writer .`
- **KEEP:** 10% of cases, we keep them unchanged
- **REPLACE:** 10% of cases, we replace them with random sentences

  `William Shakespeare is a poet . Birds can fly . He is regarded as the greatest writer .`

Experiments

Datasets

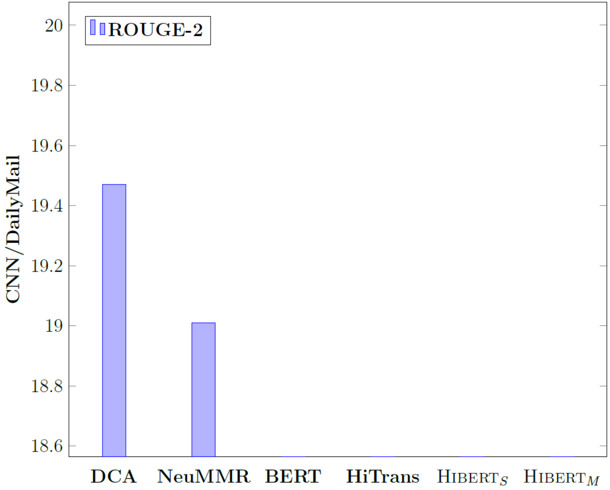
Dataset	Train	Dev	Test
Gigaword	6,626,842	13,368	—
CNNDM	287,226	13,368	11,490
NYT50	137,778	17,222	17,223

- **Gigaword**: Part of Giagaword, 2.8 billion words
  - Used for pre-training
- **CNNDM**: CNN/DailyMail Dataset (Hermann et al., 2015)
- **NYT50**: New York Times Dataset
  - remove documents whose summaries are shorter than 50 words (Durrett et al., 2016; Xu and Durrett, 2019)

Training Details

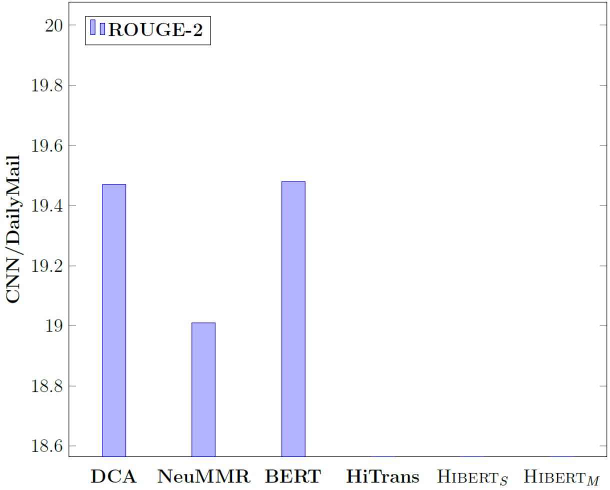
- Three-stage Training:
  - Open-domain Pre-training (Gigaword)
  - In-domain Pre-training (CNNDM or NYT50)
  - Fine-tuning on CNNDM or NYT50
- Batch Size: 256 documents; 45 epochs for open-domain, 100 to 200 epochs for in-domain pre-training
- HIBERT_S: L = 6, H = 512 and A = 8
- HIBERT_M: L = 6, H = 768 and A = 12
- around 20 hours per epoch for HIBERT_M with 8 Nvidia Tesla V100 GPUs, open domain pre-training takes around 35 days!

Automatic Evaluation: CNN/DailyMail Dataset



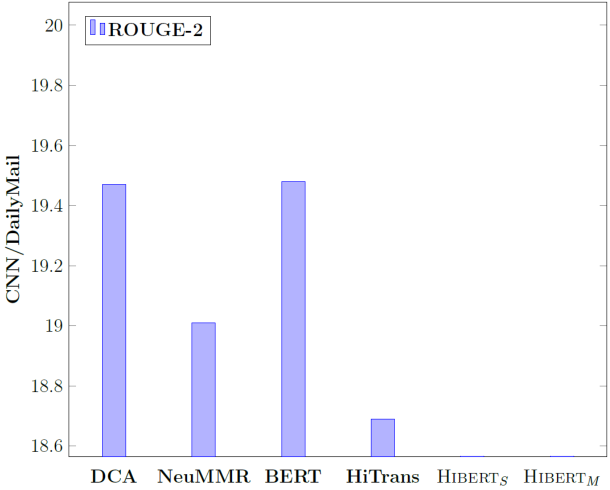
- DCA (Celikyilmaz et al., 2018), NeuMMR (Zhou et al., 2019)
- # Params: BERT (110M), HIBERT$_S$ (54.6M), HIBERT$_M$ (110M)

Automatic Evaluation: CNN/DailyMail Dataset



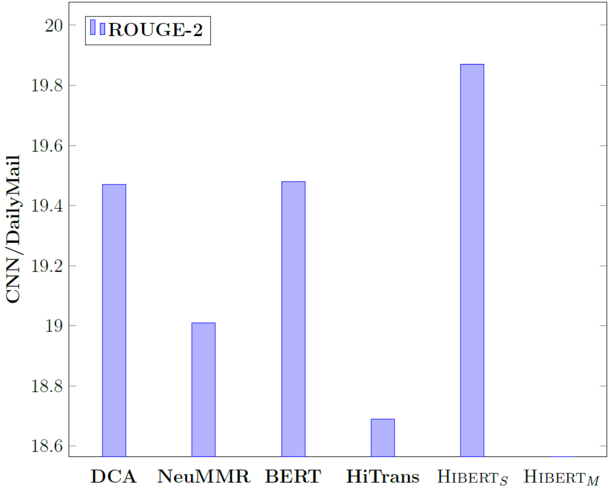
- DCA (Celikyilmaz et al., 2018), NeuMMR (Zhou et al., 2019)
- # Params: BERT (110M), HIBERT$_S$ (54.6M), HIBERT$_M$ (110M)

Automatic Evaluation: CNN/DailyMail Dataset



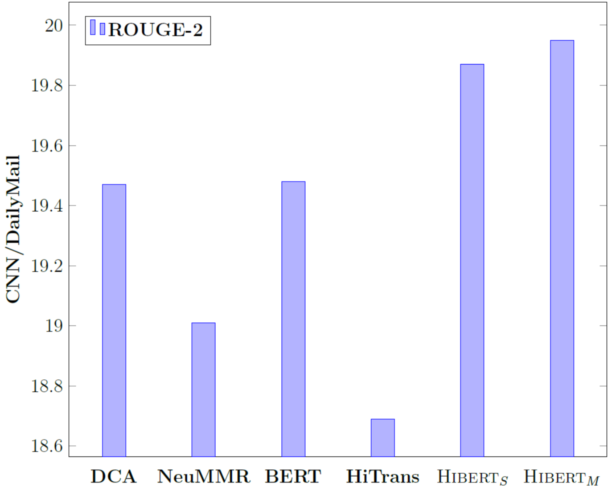
- DCA (Celikyilmaz et al., 2018), NeuMMR (Zhou et al., 2019)
- # Params: BERT (110M), HIBERT$_S$ (54.6M), HIBERT$_M$ (110M)

Automatic Evaluation: CNN/DailyMail Dataset



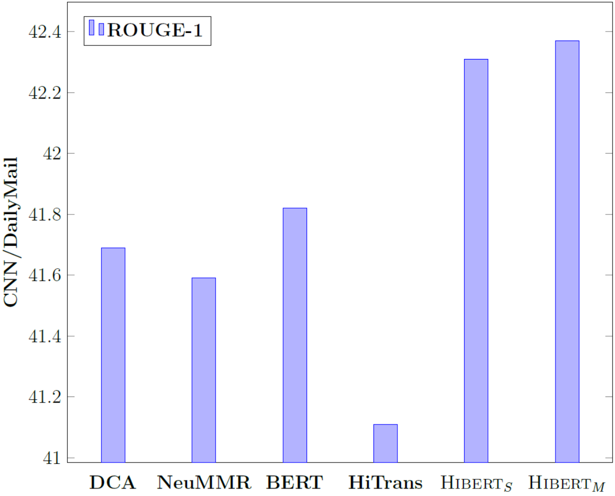
- DCA (Celikyilmaz et al., 2018), NeuMMR (Zhou et al., 2019)
- # Params: BERT (110M), HIBERT$_S$ (54.6M), HIBERT$_M$ (110M)

Automatic Evaluation: CNN/DailyMail Dataset



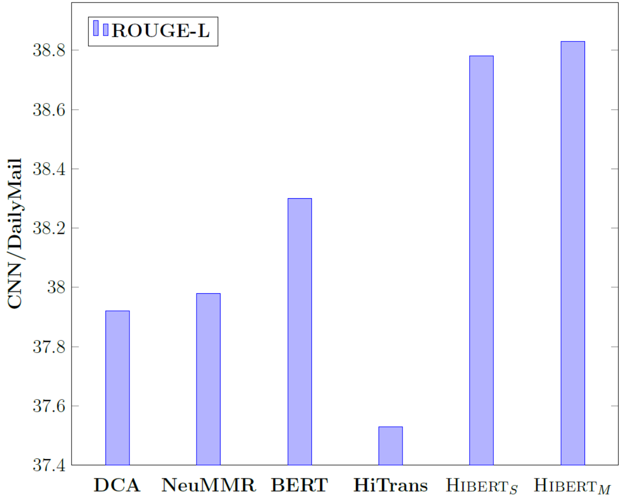
- DCA (Celikyilmaz et al., 2018), NeuMMR (Zhou et al., 2019)
- # Params: BERT (110M), HIBERT$_S$ (54.6M), HIBERT$_M$ (110M)

Automatic Evaluation: CNN/DailyMail Dataset



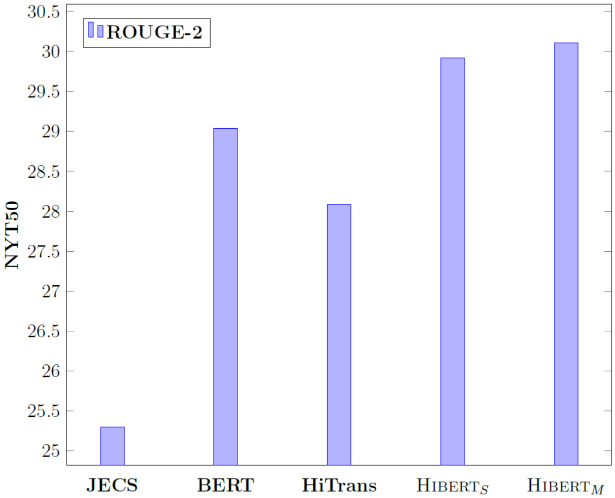
- DCA (Celikyilmaz et al., 2018), NeuMMR (Zhou et al., 2019)
- # Params: BERT (110M), HIBERT$_S$ (54.6M), HIBERT$_M$ (110M)

Automatic Evaluation: CNN/DailyMail Dataset



- DCA (Celikyilmaz et al., 2018), NeuMMR (Zhou et al., 2019)
- # Params: BERT (110M), HIBERT$_S$ (54.6M), HIBERT$_M$ (110M)

Automatic Evaluation: NYT50 Dataset



JECS (Xu and Durrett, 2019)

Human Evaluation: CNN/DailyMail Dataset

Models	1st	2nd	3rd	4th	5th	6th	MeanR
Lead3	0.03	0.18	0.15	0.30	0.30	0.03	3.75
DCA	0.08	0.15	0.18	0.20	0.15	0.23	3.88
Latent	0.05	0.33	0.28	0.20	0.13	0.00	3.03
BERT	0.13	0.37	0.32	0.15	0.03	0.00	2.58
HIBERT_M	0.30	0.35	0.25	0.10	0.00	0.00	2.15
Human	0.58	0.15	0.20	0.00	0.03	0.03	1.85

- DCA (Celikyilmaz et al., 2018); Latent (Zhang et al., 2018)
- MeanR: Mean Ranks; Lower is better
- HIBERT_M is significantly better than all models except for Human ($p < 0.05$ with student t-test)

Conclusions

- The core part of a neural extractive summarization model is the hierarchical document encoder
- We proposed a method to pre-train it on unlabeled data
- Experiments show the pre-training method is effective
- Future Work:
  - Apply HIBERT to other tasks
  - Improve architectures of HIBERT
  - New and *free* pre-training tasks