2022.08.07 중급6조

PEGASUS

Pre-training with Extracted Gap-sentences for Abstractive Summarization

2022.08.07 중급6조

PEGASUS

Pre-training with Extracted Gap-sentences for Abstractive Summarization

- **ICML 2020**
- **679회 인용** (2022.08.07 기준)
- **Jingqing Zhan[1], Yao Zhao[2], Mohammad Saleh[2], Peter J. Liu[2]**

[1] Brain Team, Google Research, Mountain View, CA, USA

[2] Data Science Institute, Imperial College London, London, UK

2022.08.07 중급6조

PEGASUS

Pre-training with Extracted Gap-sentences for Abstractive Summarization

1 Introduction

2 Pre-training Objectives

3 Pre-training Corpus & Downstream Tasks/Datasets

4 Experiments

- 4.1 Ablations on PEGASUS$_{\text{BASE}}$
- 4.2 Larger Model Results
- 4.3 Zero and Low-Resource Summarization
- 4.4 Qualitative Observations and Human Evaluation
- 4.5 Test-set Overlap with Pre-training Corpus
- 4.6 Additional PEGASUS$_{\text{LARGE}}$ Improvements

5 Conclusion

1 Introduction

[등장 배경]

- 추상적 요약만을 목적으로 pre-training된 모델의 부재

→ Gap-Sentences Generation (GSG) objective 제안

- 지난 몇년 간 새로운 추상 요약 데이터셋 구축에 대한 관심이 증가한 추세이나

  이에 대한 모델의 체계적인 평가 연구 부족

→ **다양한 도메인의 12개 요약 데이터셋으로 성능 평가**

1 Introduction

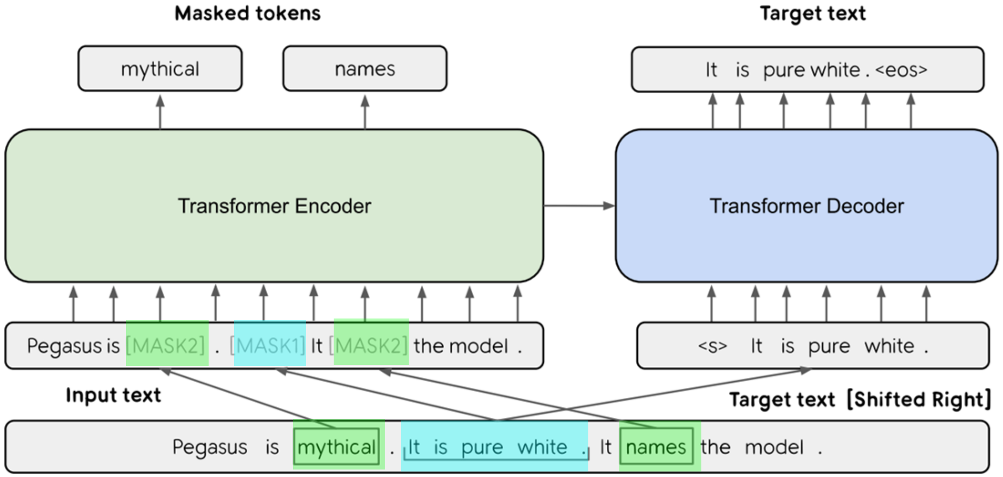
[PEGASUS]

- 새로운 self-supervised objective 'GSG'로 대규모 텍스트 말뭉치에 대해 사전학습한 Transformer기반 인코더-디코더 모델

Pre-training with
Extracted
Gap-sentences for
Abstractive
SUmmarization
Sequence-to-sequence models

[모델 구조]

- Transformer based encoder-decoder model
- pre-training objective: GSG



Masked tokens
mythical
names
Target text
It is pure white . <eos>
Transformer Encoder
Transformer Decoder
Pegasus is [MASK2] . [MASK1] It [MASK2] the model .
<s> It is pure white .
Input text
Target text [Shifted Right]
Pegasus is mythical . It is pure white . It names the model .

< MLM과 GSG를 pre-training objective로 사용한 경우 PEGASUS 기본 아키텍처 >

Masked Language Model *Gap-Sentences Generation*

※ 최종적으로는 GSG만 채택함 ※

2 Pre-training Objectives

[Gap Sentences Generation (GSG)]

- 추상적 요약을 위해 제안된 새로운 self-supervised pre-training objective
- 문장단위로 마스킹하고 나머지 문장들을 통해 마스킹된 문장을 예측하는 방식으로 학습
- downstream task와 더 유사한 pre-training objective일수록
  fine-tuning 성능이 더 빠르고 좋을 것이라는 가정에서 나오게 되었다.

[GSG 동작 과정]

1. GSR(gap sentences ratio)에 따라 문장들을 선택 및 마스킹한다.
2. 선택된 gap sentence들을 연결하여 하나의 pseudo-요약문으로 만든다.
3. gap sentence들의 각 위치에 [MASK1]토큰을 넣은 후 모델에 넣는다.

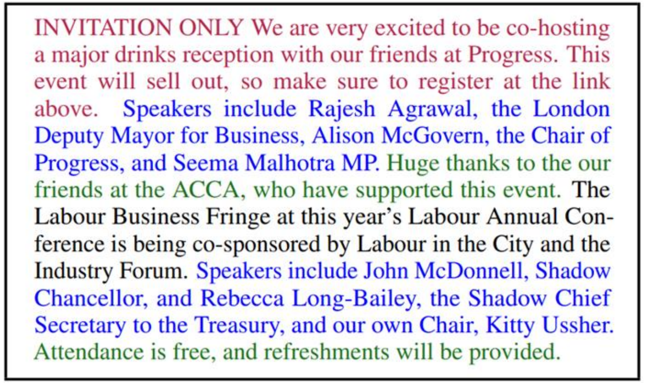
TRANSFORMER

※ **GSR(gap sentences ratio)**: 문서 내 전체 문장 개수 대비 선택된 gap sentence의 개수 비율.
최종적으로 30% 채택함.

2 Pre-training Objectives

[Gap Sentence Selection Methods]

- Random: 랜덤하게 m개의 문장을 선택
- Lead : 문서의 처음 m개 문장을 선택
- Principal : 각 문장의 중요도를 매겨서 상위 m개 문장을 선택 (ROUGE1-F1 Score)
  - 옵션1. Ind or Seq: 문장을 순서와 관계없이 선택 or 연속적으로 선택
  - 옵션2. Uniq or Orig: n-gram을 집합으로 취급 or 중복 허용



INVITATION ONLY We are very excited to be co-hosting a major drinks reception with our friends at Progress. This event will sell out, so make sure to register at the link above. Speakers include Rajesh Agrawal, the London Deputy Mayor for Business, Alison McGovern, the Chair of Progress, and Seema Malhotra MP. Huge thanks to the our friends at the ACCA, who have supported this event. The Labour Business Fringe at this year's Labour Annual Conference is being co-sponsored by Labour in the City and the Industry Forum. Speakers include John McDonnell, Shadow Chancellor, and Rebecca Long-Bailey, the Shadow Chief Secretary to the Treasury, and our own Chair, Kitty Ussher. Attendance is free, and refreshments will be provided.

Figure 2: An example of sentences (from the C4 corpus) selected by Random, Lead and Ind-Orig respectively. Best viewed in color.

3 Pre-training Corpus & Downstream Tasks/Datasets

[Pre-training Corpus]

- C4: Colossal Clean Crawled Corpus. T5연구진이 웹 크롤링으로 수집한 350M개 웹페이지의 텍스트 데이터셋. (750GB)
- HugeNews: 새로 수집한 2013-2019년도 뉴스 기사 데이터셋. (3.8TB) 뉴스 출간물과 같은 높은 퀄리티부터 블로그나 고등학교 신문처럼 낮은 퀄리티까지 모두 아우른다.

[Downstream Datasets]

- 재현 가능한 코드 제공을 위해 public 데이터셋인 TensorFlow Summarization Datasets을 사용하였다.
- XSum, CNN/DailyMail, NEWSROOM, Multi-News, Gigaword, WikiHow, Reddit TIFU, BIGPATENT, arXiv, PubMed, AESLC, BillSum
- 뉴스 기사, 특허, 과학, 법률, 이메일 등 다양한 주제의 데이터셋

https://www.tensorflow.org/datasets/catalog/overview#summarization

4 Experiments

초기 실험

- 시간과 자원을 절약하기 위해 사이즈를 줄인 PEGASUS-Base를 이용
- 12개의 데이터셋 중 4개만을 이용 (XSum, CNN/DailyMail, WikiHow, Reddit TIFU)

	Encoder, Decoder Layer 개수	Hidden Size	Feed-Forward Layer 개수	Self-Attention Head 개수	Batch Size	파라미터 수
PEGASUS-Base	12	768	3072	12	256	223M
PEGASUS-Large	16	1024	4096	16	8192	568M

Optimizer

- Adafactor : 메모리 사용량을 줄일 수 있는 Adam기반의 Optimizer

Dropout rate

- 0.1

4.1 Ablations on PEGASUS$_{BASE}$

PEGASUS_Base의 평가항목

- Pre-training Corpus
- Pre-training 방식
- 단어장 사이즈

Pre-training Corpus

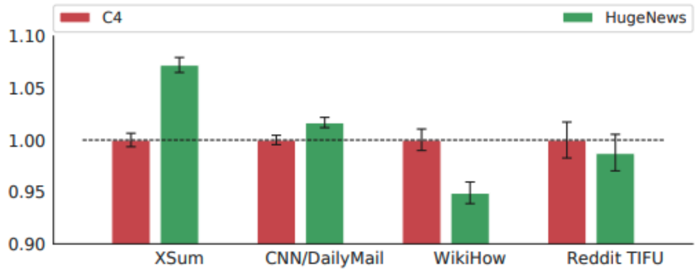


Figure 3: Effect of pre-training corpus. PEGASUS$_{BASE}$ pre-trained on C4 (350M Web-pages) and HugeNews (1.5B news-like documents).

- 뉴스가 아닌 데이터셋 : C4가 효율적
- 뉴스 DownStream 데이터셋 : HugeNews가 효율적

4.1 Ablations on PEGASUS_{BASE}

PEGASUS_Base의 평가항목

- Pre-training Corpus
- Pre-training 방식
- 단어장 사이즈

→ C4만을 이용해서 실험

Pre-training Corpus

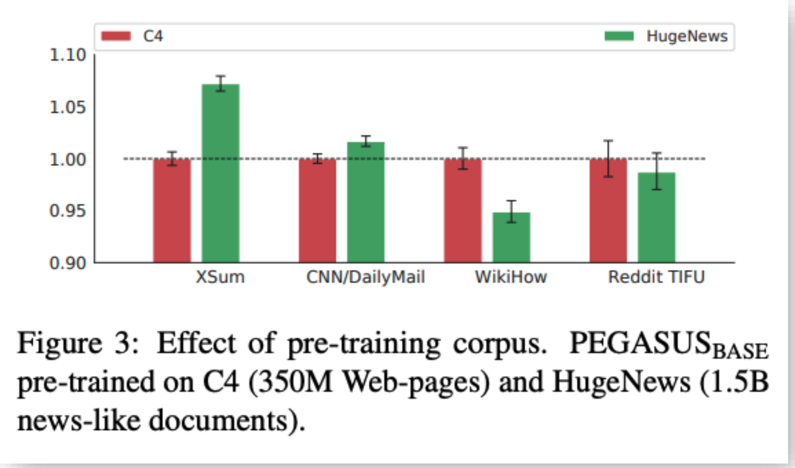
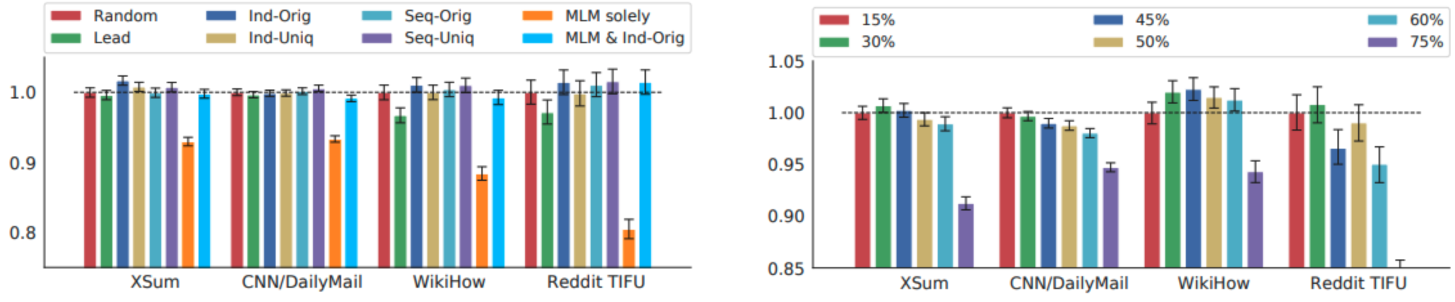


Figure 3: Effect of pre-training corpus. $\text{PEGASUS}_{\text{BASE}}$ pre-trained on C4 (350M Web-pages) and HugeNews (1.5B news-like documents).

- 뉴스가 아닌 데이터셋 : C4가 효율적
- 뉴스 DownStream 데이터셋 : HugeNews가 효율적

4.1 Ablations on PEGASUS$_{BASE}$

Pre-training 방식



(a) Effect of pre-training objectives (30% GSR).

(b) Effect of gap sentences ratio with GSG (Ind-Orig).

Figure 4: Effect of pre-training settings with PEGASUS$_{\text{BASE}}$ pre-trained on C4.

- 30%의 GSR에서 실험
- GSG : Ind-Orig > Seq-Uniq 순으로 성능이 좋음
- MLM : 단독으로 사용했을 경우 성능이 더 낮았고, GSG와 혼합한 방식도 GSG의 Random 방식과 비슷한 성능을 보임

- Ind-Orig 방식에서 실험
- 공평한 실험을 위해 데이터셋을 최대 400단어를 갖도록 자름
- 50%이하일 때 대체적으로 좋은 성능을 보임
- CNN/DailyMail : 15%일때 성능이 좋음
- XSum,Reddit TIFU : 30%일때 성능이 좋음
- WikiHow : 45%일때 성능이 좋음

4.1 Ablations on PEGASUS$_{BASE}$

단어장 사이즈

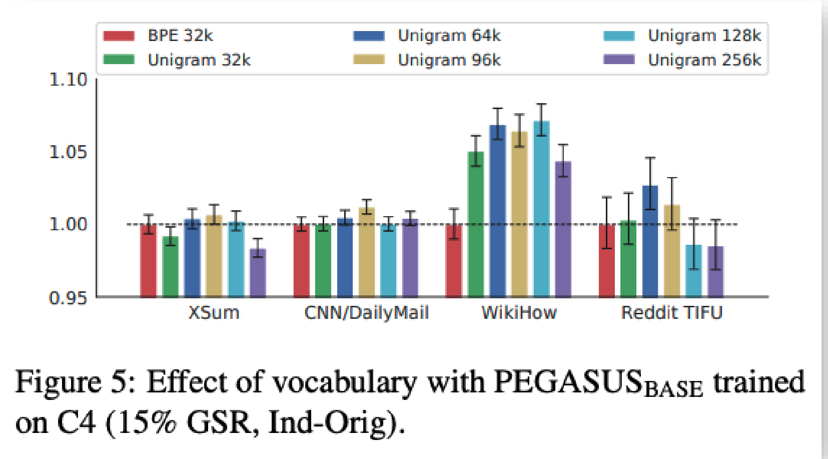


Figure 5: Effect of vocabulary with PEGASUS$_{\text{BASE}}$ trained on C4 (15% GSR, Ind-Orig).

- 두가지 토크나이저를 사용 (Byte-pair-encoding = BPE, SentencePiece Unigram = Unigram)
- XSum, CNN/DailyMail : Unigram 96k일때 성능이 좋음
- WikiHow : Unigram 128k일때 성능이 좋음
- Reddit TIFU : Unigram 96k일때 성능이 좋음

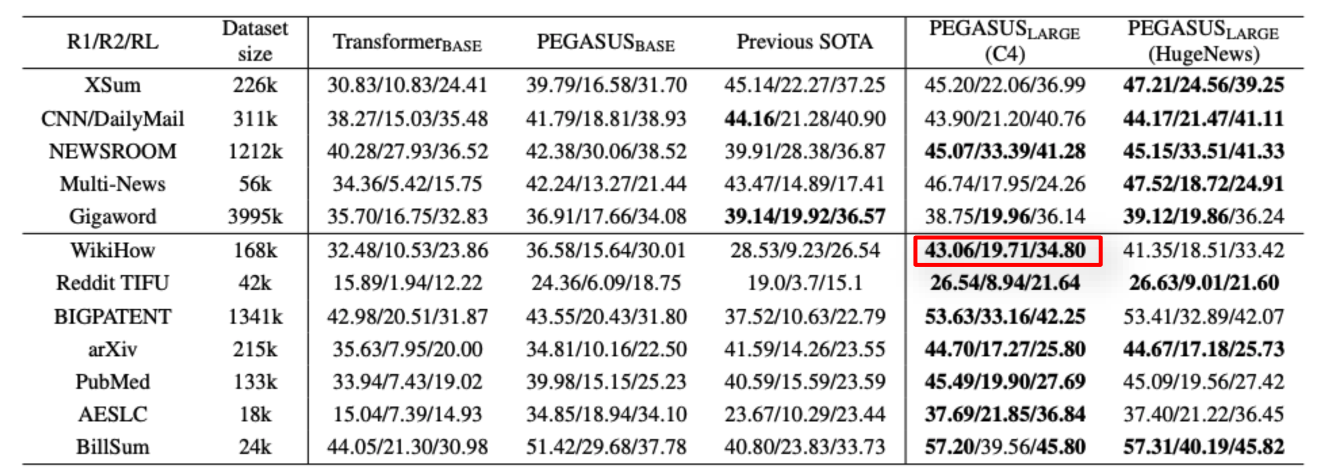
4.2 Larger Model Results

PEGASUS_Base 실험 후 적용 사항

- Pre-training 방식 : GSG(Ind-Orig)
- GSR : 45% (30%와 비슷한 성능을 내기 위해 선택)
- 단어장 사이즈 : Unigram 96k

	Encoder, Decoder Layer 개수	Hidden Size	Feed-Forward Layer 개수	Self-Attention Head 개수	Batch Size	파라미터 수	Optimizer	Dropout Rate
PEGASUS-Large	16	1024	4096	16	8192	568M	Adafactor	0.1

4.2 Larger Model Results



R1/R2/RL	Dataset size	$\text{Transformer}_{\text{BASE}}$	$\text{PEGASUS}_{\text{BASE}}$	Previous SOTA	$\text{PEGASUS}_{\text{LARGE}}$ (C4)	$\text{PEGASUS}_{\text{LARGE}}$ (HugeNews)
XSum	226k	30.83/10.83/24.41	39.79/16.58/31.70	45.14/22.27/37.25	45.20/22.06/36.99	**47.21/24.56/39.25**
CNN/DailyMail	311k	38.27/15.03/35.48	41.79/18.81/38.93	**44.16**/21.28/40.90	43.90/21.20/40.76	**44.17/21.47/41.11**
NEWSROOM	1212k	40.28/27.93/36.52	42.38/30.06/38.52	39.91/28.38/36.87	**45.07/33.39/41.28**	**45.15/33.51/41.33**
Multi-News	56k	34.36/5.42/15.75	42.24/13.27/21.44	43.47/14.89/17.41	46.74/17.95/24.26	**47.52/18.72/24.91**
Gigaword	3995k	35.70/16.75/32.83	36.91/17.66/34.08	**39.14/19.92/36.57**	38.75/**19.96**/36.14	**39.12/19.86**/36.24
WikiHow	168k	32.48/10.53/23.86	36.58/15.64/30.01	28.53/9.23/26.54	**43.06/19.71/34.80**	41.35/18.51/33.42
Reddit TIFU	42k	15.89/1.94/12.22	24.36/6.09/18.75	19.0/3.7/15.1	**26.54/8.94/21.64**	**26.63/9.01/21.60**
BIGPATENT	1341k	42.98/20.51/31.87	43.55/20.43/31.80	37.52/10.63/22.79	**53.63/33.16/42.25**	53.41/32.89/42.07
arXiv	215k	35.63/7.95/20.00	34.81/10.16/22.50	41.59/14.26/23.55	**44.70/17.27/25.80**	**44.67/17.18/25.73**
PubMed	133k	33.94/7.43/19.02	39.98/15.15/25.23	40.59/15.59/23.59	**45.49/19.90/27.69**	45.09/19.56/27.42
AESLC	18k	15.04/7.39/14.93	34.85/18.94/34.10	23.67/10.29/23.44	**37.69/21.85/36.84**	37.40/21.22/36.45
BillSum	24k	44.05/21.30/30.98	51.42/29.68/37.78	40.80/23.83/33.73	**57.20**/39.56/**45.80**	**57.31/40.19/45.82**

- PEGASUS_Base도 대부분의 Downstream에서 현재 SOTA에 비해 좋은 성능을 내고 있지만, PEGASUS_Large의 경우 모든 Downstream에서 SOTA보다 좋은 성능을 내고 있음
- 대부분 Downstream은 HugeNews로 사전학습한 모델이 더 높은 성능을 보임
- 단, WikiHow의 경우 C4로 사전학습한 모델이 더 높은 성능을 보임

4.3 Zero and Low-Resource Summarization

▶ 실험 Setting

- Pegasus_large model (HugeNews)
  - ○ 12개 dataset에서 10^k (k=1,2,3,4) training examples을 추출함
  - ○ 2000 steps with batch size 256, learning rate=0.00005
- Transformer-base model은 full supervised dataset으로 fine-tuning됨

▶ 실험 결과 (Figure 6)

- Pegasus-large model은 fully-supervised dataset으로 학습된 Transformer-base model과 비슷한 quality의 summaries를 생성함
- *Multi-News, WikiHow, Reddit TIFU, BigPatent, AESLC, BillSum* : 1000 examples로 학습한 Pegasus-large model이 이전 SOTA보다 ROUGE score가 높음

▼ *Figure 6 : Fine-tuning with limited supervised examples*

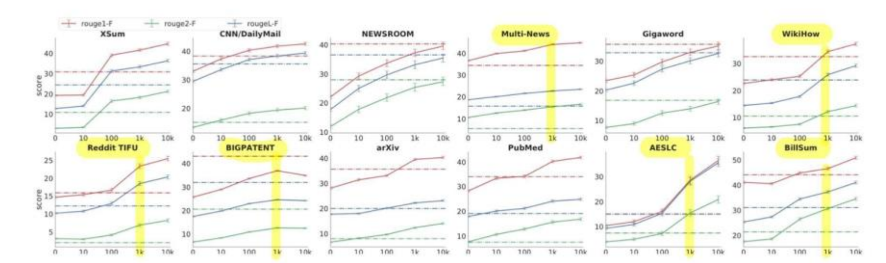


Figure 6: Fine-tuning with limited supervised examples. The solid lines are PEGASUS$_{LARGE}$ fine-tuned on 0 (zero shot), 10, 100, 1k,10k examples. The dashed lines are Transformer$_{BASE}$ models, equivalent in capacity as PEGASUS$_{BASE}$ and trained using the full supervised datasets, but with no pre-training. All numbers are reported in Appendix E.

▼ *Appendix E*

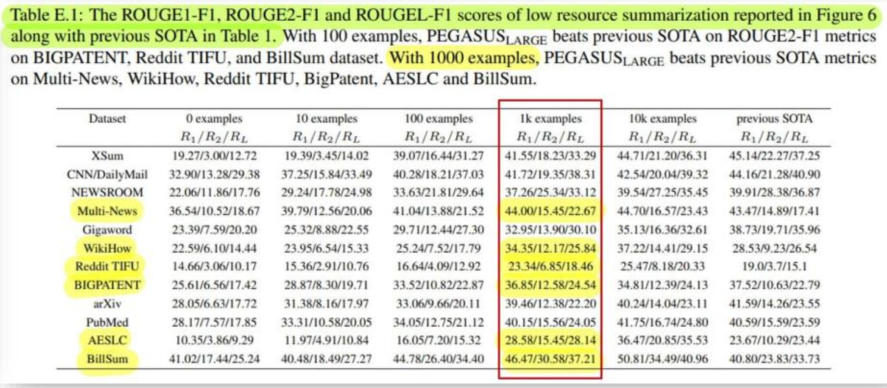


Table E.1: The ROUGE1-F1, ROUGE2-F1 and ROUGEL-F1 scores of low resource summarization reported in Figure 6 along with previous SOTA in Table 1. With 100 examples, PEGASUS$_{LARGE}$ beats previous SOTA on ROUGE2-F1 metrics on BIGPATENT, Reddit TIFU, and BillSum dataset. With 1000 examples, PEGASUS$_{LARGE}$ beats previous SOTA metrics on Multi-News, WikiHow, Reddit TIFU, BigPatent, AESLC and BillSum.

Dataset	0 examples	10 examples	100 examples	1k examples	10k examples	previous SOTA
	$R_1/R_2/R_L$	$R_1/R_2/R_L$	$R_1/R_2/R_L$	$R_1/R_2/R_L$	$R_1/R_2/R_L$	$R_1/R_2/R_L$
XSum	19.27/3.00/12.72	19.39/3.45/14.02	39.07/16.44/31.27	41.55/18.23/33.29	44.71/21.20/36.31	45.14/22.27/37.25
CNN/DailyMail	32.90/13.28/29.38	37.25/15.84/33.49	40.28/18.21/37.03	41.72/19.35/38.31	42.54/20.04/39.32	44.16/21.28/40.90
NEWSROOM	22.06/11.86/17.76	29.24/17.78/24.98	33.63/21.81/29.64	37.26/25.34/33.12	39.54/27.25/35.45	39.91/28.38/36.87
Multi-News	36.54/10.52/18.67	39.79/12.56/20.06	41.04/13.88/21.52	44.00/15.45/22.67	44.70/16.57/23.43	43.47/14.89/17.41
Gigaword	23.39/7.59/20.20	25.32/8.88/22.55	29.71/12.44/27.30	32.95/13.90/30.10	35.13/16.36/32.61	38.73/19.71/35.96
WikiHow	22.59/6.10/14.44	23.95/6.54/15.33	25.24/7.52/17.79	34.35/12.17/25.84	37.22/14.41/29.15	28.53/9.23/26.54
Reddit TIFU	14.66/3.06/10.17	15.36/2.91/10.76	16.64/4.09/12.92	23.34/6.85/18.46	25.47/8.18/20.33	19.0/3.7/15.1
BIGPATENT	25.61/6.56/17.42	28.87/8.30/19.71	33.52/10.82/22.87	36.85/12.58/24.54	34.81/12.39/24.13	37.52/10.63/22.79
arXiv	28.05/6.63/17.72	31.38/8.16/17.97	33.06/9.66/20.11	39.46/12.38/22.20	40.24/14.04/23.11	41.59/14.26/23.55
PubMed	28.17/7.57/17.85	33.31/10.58/20.05	34.05/12.75/21.12	40.15/15.56/24.05	41.75/16.74/24.80	40.59/15.59/23.59
AESLC	10.35/3.86/9.29	11.97/4.91/10.84	16.05/7.20/15.32	28.58/15.45/28.14	36.47/20.85/35.53	23.67/10.29/23.44
BillSum	41.02/17.44/25.24	40.48/18.49/27.27	44.78/26.40/34.40	46.47/30.58/37.21	50.81/34.49/40.96	40.80/23.83/33.73

4.4 Qualitative Observations and Human Evaluation

▶ Human Evaluation

- 평가방법 : Amazon Mechanical Turk site를 통해, human evaluation을 진행. 1점 (Poor) ~ 5점 (Great)까지 점수를 매기도록 진행함
- 평가 데이터셋 : *XSum, CNN/DM, Reddit TIFU*

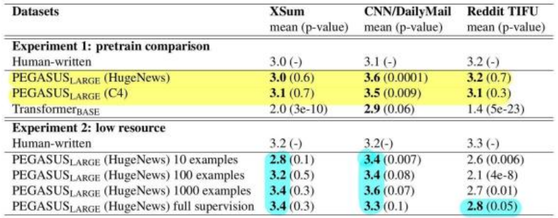
1) 1번째 실험 (Experiment 1: pretrain comparison)

- 실험 방법 : *Pegasus-large (HugeNews), Pegasus-large (C4), Transformer-base model*이 생성한 요약문과 reference summaries를 비교함
- 실험 결과 : Human evaluation에서, Pegasus-large (HugeNews)와 Pegasus-large (C4) 모두 Human-written summary (3.0 XSum, 3.1 CNNDM, 3.2 Reddit TIFU)와 비슷하거나 높은 score를 기록

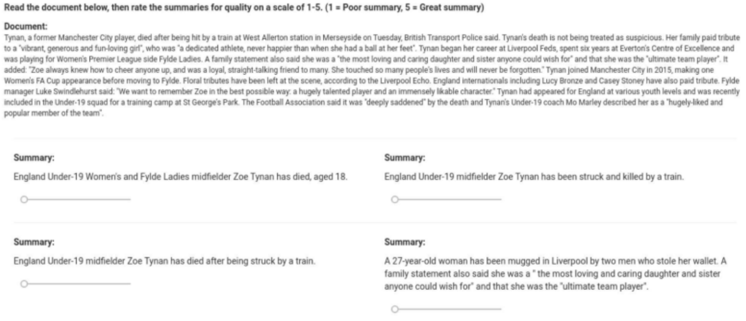
1) 2번째 실험 (Experiment 2: low resource)

- 실험 방법 : *Pegasus-large (HugeNews)*는 10, 100, 1000 examples로 fine-tuning한 후, 각 model이 생성한 요약문과 reference summaries를 비교함
- 실험 결과 :
  - XSum, CNNDM dataset) 10 ~ 1000 examples로만 학습한 Pegasus-large model이여도 Human-written과 유사하거나 높은 score를 기록
  - 하지만, Reddit TIFU에서는 full supervision dataset으로 학습한 model만 유사한 score를 기록하였는데, dataset의 특성(diverse writing style)때문이라 추측

Table 3: Human evaluation side-by-side results on Likert (1-5) scale (higher is better). Scores are bolded if they are not worse than human-level performance by $p < 0.01$.



Datasets	XSum mean (p-value)	CNN/DailyMail mean (p-value)	Reddit TIFU mean (p-value)
Experiment 1: pretrain comparison			
Human-written	3.0 (-)	3.1 (-)	3.2 (-)
PEGASUS$_{\text{LARGE}}$ (HugeNews)	**3.0** (0.6)	**3.6** (0.0001)	**3.2** (0.7)
PEGASUS$_{\text{LARGE}}$ (C4)	**3.1** (0.7)	**3.5** (0.009)	**3.1** (0.3)
Transformer$_{\text{BASE}}$	2.0 (3e-10)	**2.9** (0.06)	1.4 (5e-23)
Experiment 2: low resource			
Human-written	3.2 (-)	3.2(-)	3.3 (-)
PEGASUS$_{\text{LARGE}}$ (HugeNews) 10 examples	**2.8** (0.1)	**3.4** (0.007)	2.6 (0.006)
PEGASUS$_{\text{LARGE}}$ (HugeNews) 100 examples	**3.2** (0.5)	**3.4** (0.08)	2.1 (4e-8)
PEGASUS$_{\text{LARGE}}$ (HugeNews) 1000 examples	**3.4** (0.3)	**3.6** (0.07)	2.7 (0.01)
PEGASUS$_{\text{LARGE}}$ (HugeNews) full supervision	**3.4** (0.3)	**3.3** (0.1)	**2.8** (0.05)



Read the document below, then rate the summaries for quality on a scale of 1-5. (1 = Poor summary, 5 = Great summary)

Document:
Tynan, a former Manchester City player, died after being hit by a train at West Allerton station in Merseyside on Tuesday, British Transport Police said. Tynan's death is not being treated as suspicious. Her family paid tribute to a "vibrant, generous and fun-loving girl", who was "a dedicated athlete, never happier than when she had a ball at her feet". Tynan began her career at Liverpool Feds, spent six years at Everton's Centre of Excellence and was playing for Women's Premier League side Fylde Ladies. A family statement also said she was a "the most loving and caring daughter and sister anyone could wish for" and that she was the "ultimate team player". It added: "Zoe always knew how to cheer anyone up, and was a loyal, straight-talking friend to many. She touched so many people's lives and will never be forgotten." Tynan joined Manchester City in 2015, making one Women's FA Cup appearance before moving to Fylde. Floral tributes have been left at the scene, according to the Liverpool Echo. England internationals including Lucy Bronze and Casey Stoney have also paid tribute. Fylde manager Luke Swindlehurst said: "We want to remember Zoe in the best possible way: a hugely talented player and an immensely likable character." Tynan had appeared for England at various youth levels and was recently included in the Under-19 squad for a training camp at St George's Park. The Football Association said it was "deeply saddened" by the death and Tynan's Under-19 coach Mo Marley described her as a "hugely-liked and popular member of the team".

Summary:
England Under-19 Women's and Fylde Ladies midfielder Zoe Tynan has died, aged 18.

Summary:
England Under-19 midfielder Zoe Tynan has been struck and killed by a train.

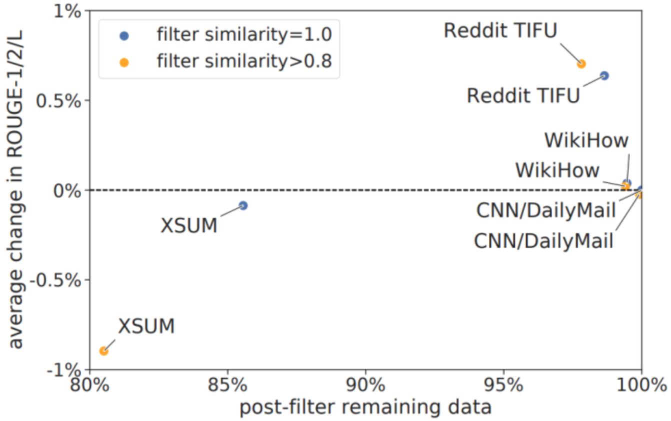
Summary:
England Under-19 midfielder Zoe Tynan has died after being struck by a train.

Summary:
A 27-year-old woman has been mugged in Liverpool by two men who stole her wallet. A family statement also said she was a " the most loving and caring daughter and sister anyone could wish for" and that she was the "ultimate team player".

Figure F.1: A screenshot of the Amazon MTurk HIIT.

4.5 Test-set Overlap with Pre-training Corpus

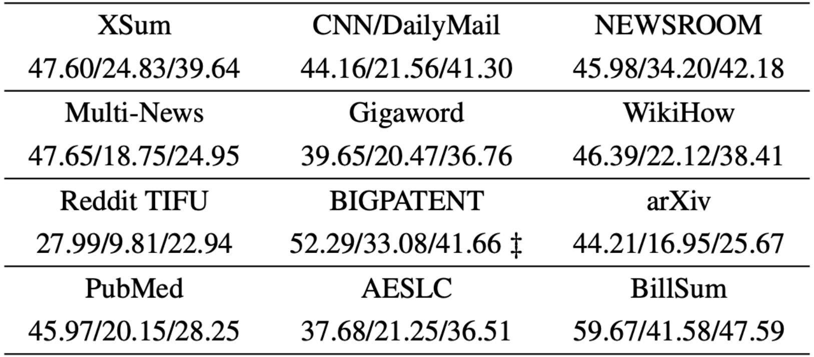
- 사전 훈련하는 코퍼스와 테스트 셋에 겹치는 정도를 평가
  - 사전 훈련하는 코퍼스가 인터넷에서 수집되었기 때문에 downstream task와 겹치는 부분이 있을 수 있음
  - 사전 훈련된 모델이 이를 기억하고 downstream test에서 더 좋은 성능을 냈는지에 대한 부분도 평가
- test set과 pre-training corpus 사이의 유사도를 ROUGE-2 recall 점수로 측정 (2-grams)
  - threshold를 넘는 sample은 필터링함



- 그림은 사전 훈련 코퍼스 C4와 XSum, CNN/Dailymail, Reddit TIFU, WikiHow의 overlap을 1.0, 8.0 threshold 값으로 평가
- XSUM 데이터셋의 overlap이 가장 높았음
- 필터링한 데이터셋으로 다시 test set 점수를 구했을 때 ROUGE 점수차이는 1% 정도 밖에 나지 않았음
- 또한 overlap sample에 대해 모델이 생성한 요약문과 사람이 만든 요약문은 다른 형태를 보였음. memorization 은 거의 발생하지 않았음

4.6 Additional PEGASUS$_{LARGE}$ Improvements

- PEGASUS$_{\textit{LARGE (mixed, stochastic)}}$ 모델이 downstream task에서 가장 좋은 성능을 보였음
- (1) 해당 모델은 C4, HugeNews를 합해서 사전 훈련됨
- (2) gap 문장 비율을 균일하게 15-45% 사이로 동적으로 골랐음
- (3) 중요 문장은 점수에 대해 20% 균일한 노이즈로 stochastically sample
- (4) 150만 스텝으로 사전 훈련됨 (느리게 수렴했기 때문에)
- (5) newline character를 인코딩하도록 SentencePiece 토크나이저를 업데이트



XSum	CNN/DailyMail	NEWSROOM
47.60/24.83/39.64	44.16/21.56/41.30	45.98/34.20/42.18
Multi-News	Gigaword	WikiHow
47.65/18.75/24.95	39.65/20.47/36.76	46.39/22.12/38.41
Reddit TIFU	BIGPATENT	arXiv
27.99/9.81/22.94	52.29/33.08/41.66 ‡	44.21/16.95/25.67
PubMed	AESLC	BillSum
45.97/20.15/28.25	37.68/21.25/36.51	59.67/41.58/47.59

5 Conclusion

- PEGASUS 모델은 sequence-to-sequence 모델이며, 생성 요약을 위해 gap-sentences generation을 사전훈련의 목적으로 설정
- principle sentence selection이 gap-sentence selection 중 가장 최적의 방법론이었음을 밝힘
- 사전훈련 코퍼스, gap-sentences 비율, 단어장 크기 등의 효과를 보여줌
- 가장 좋은 configuration으로 12개 downstream datasets에서 SOTA 달성
- 모델은 또한 처음 보는 요약 데이터셋에도 빠르게 적응함. 1000개 예시만 가지고도 좋은 결과를 보임
- 사람의 평가를 통해 모델의 요약문이 사람 수준을 달성함을 보임