Liputan6: A Large-scale Indonesian Dataset for Text Summarization

Fajri Koto, Jey Han Lau, and Timothy Baldwin
The University of Melbourne

The 1st AACL and the 10th IJCNLP
2020

THE UNIVERSITY OF MELBOURNE

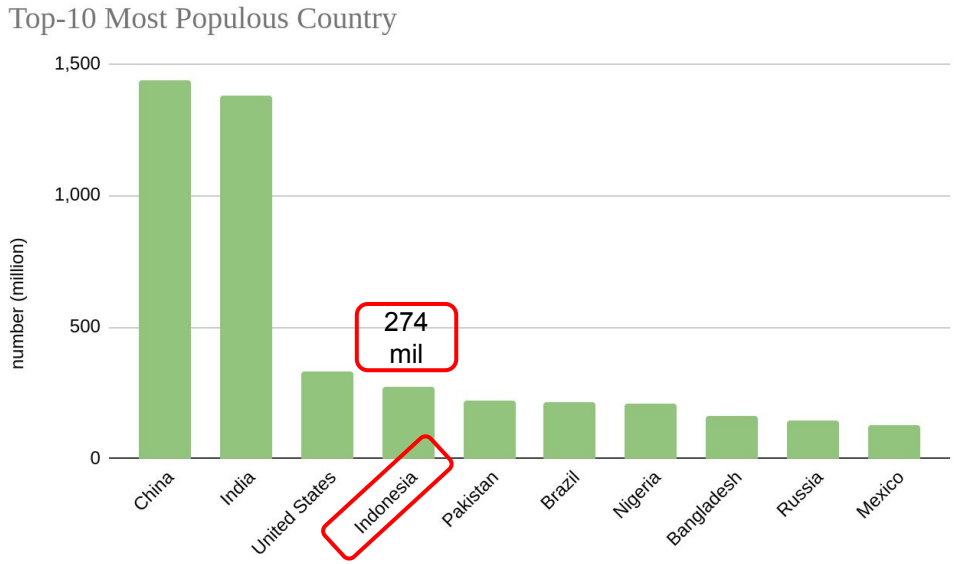
ffajri@student.unimelb.edu.au

1

Introduction

A. Introduction

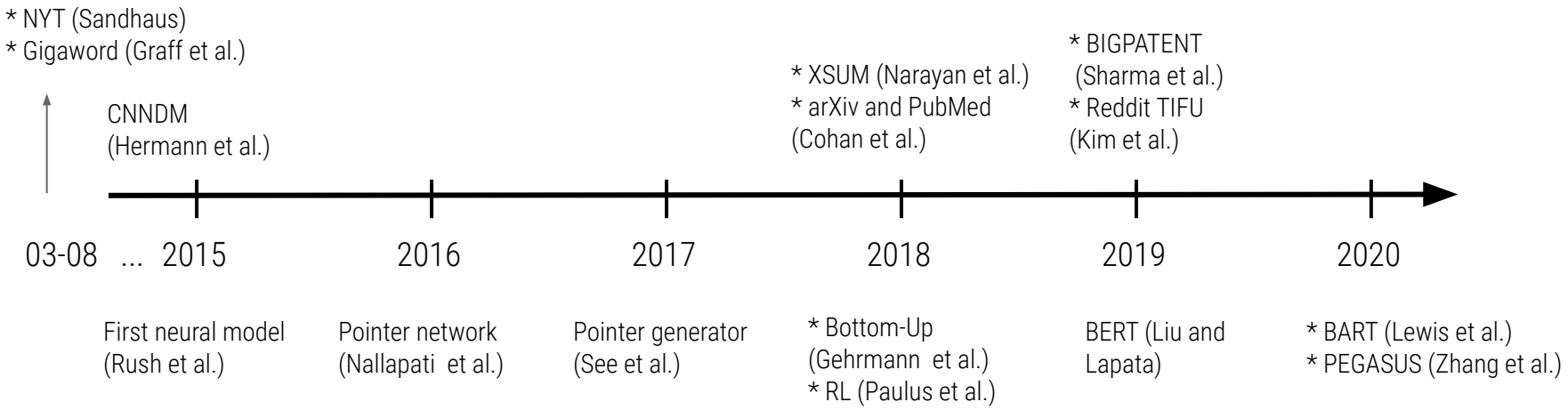
Indonesia



https://www.worldometers.info/

B. Related Works

Neural Summarization in English



B. Related Works

Neural Summarization in non-English

- Chinese (LCSTS) by Hu et al., 2015
  → 2 million Chinese short texts
  → from local microblogging website Sina Weibo

- Spanish (ES-News) by Gonzalez et al., 2019
  → 270k pairs news article
  → Not publicly available

B. Related Works

Summarization in Indonesian Language

- Primarily extractive and use small dataset.
  → Gunawan et al. (2017) used 3,075 news
  → Najibullah (2015) used 100 news

- Small publicly available corpus for abstractive model
  → Koto (2016) create chat summarization dataset (3 annotators, 300 chat logs)
  → Kurniawan and Louvan (2018) release IndoSum (19K news)

2

Dataset Construction

Gelar Operasi Yustisi, Polisi Tindak 7,5 Juta Pelanggar Protokol Covid-19

Nanda Perdana Putra
19 Okt 2020, 17:04 WIB

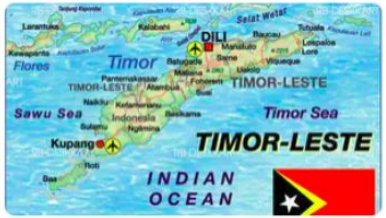
Petugas gabungan menggelar Operasi Yustisi Protokol COVID-19 di Jati Padang, Jakarta Selatan, Kamis (17/9/2020). Operasi itu untuk menegakan penerapan protokol kesehatan, terutama dalam penggunaan masker guna menekan penyebaran virus corona. (merdeka.com/Arie Basuki)

Liputan6.com, Jakarta Polisi terus menegakkan disiplin protokol kesehatan pencegahan

TOPIK POPULER
POLLYCARPUS MENINGGAL
KECELAKAAN
COVID-19
NAPI KABUR
GEMPA HARI INI

JADWAL ACARA HARI INI
SCTV INDOSIAR

SEJARAH HARI INI



Timor Timur, Provinsi Indonesia yang

A. Data Collection

https://www.liputan6.com

- We harvest this public data in 10 years time range, between October 2000 and October 2010.
- In the HTML page, the summary is located in variable `shortDescription`.
- 215,827 pairs
- Comparison with IndoSum:

Dataset	#Doc			Article			Summary		
	Train	Dev	Test	μ(Word)	μ(Sent)	#Vocab	μ(Word)	μ(Sent)	#Vocab
IndoSum	14,252	750	3,762	347.23	18.37	117K	68.09	3.47	53K
Liputan6	193,883	10,972	10,972	232.91	12.60	311K	30.43	2.09	100K

B. How abstractive is Liputan6?

https://www.liputan6.com

Dataset	LEAD-N			% of Novel n-grams			
	R1	R2	RL	1	2	3	4
IndoSum	65.6	58.9	64.8	3.1	10.8	16.2	20.3
Liputan6	41.2	27.1	38.7	12.9	41.6	57.6	66.9

C. Xtreme Dev and Test Set

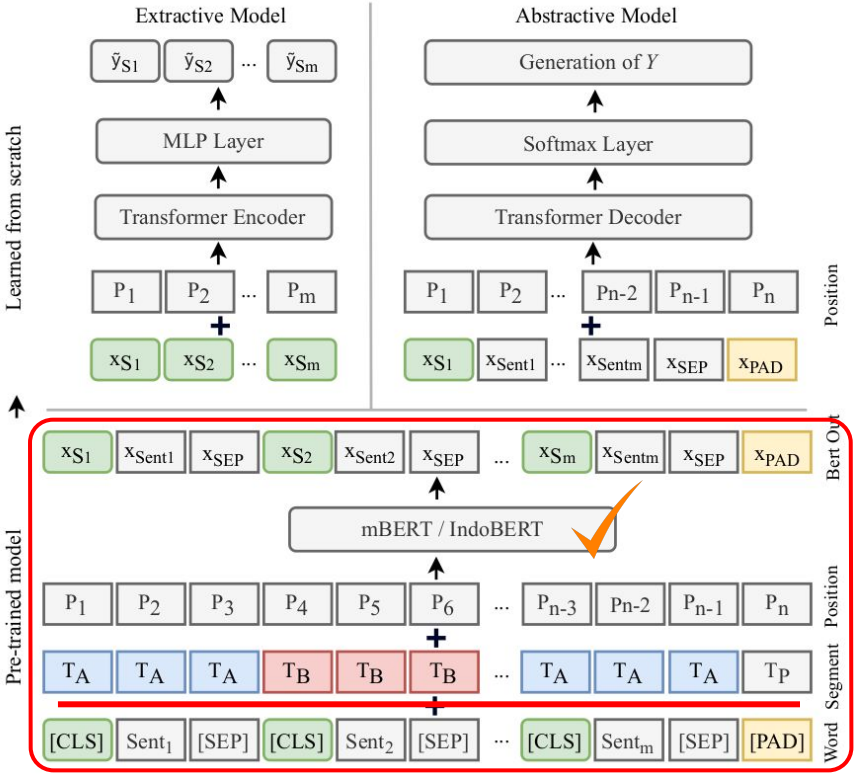
Xtreme variant provides a more abstractive summary.

Variant	#Doc			% of Novel n-grams			
	Train	Dev	Test	1	2	3	4
Canonical	193,883	10,972	10,972	16.2	52.5	71.8	82.4
Xtreme	193,883	4,948	3,862	22.2	66.7	87.5	96.6

3

Summarization Model

mBERT and IndoBERT for Summarization Model



IndoBERT

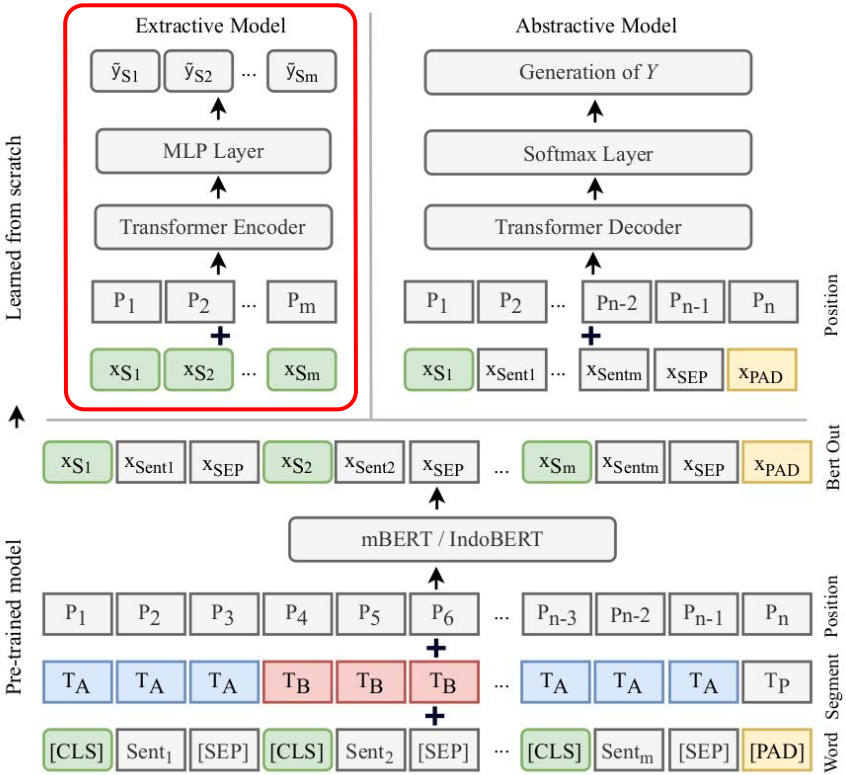
→ Data:

- ❏ Indonesian Wikipedia (74M words)
- ❏ news articles from Kompas, 10 Tempo, 11 (Tala et al., 2003) and Liputan6 (55M words)
- ❏ Indonesian Web Corpus (Medved and Suchomel, 2017) (90M words)

→ Model: bert-base-uncased
→ 1,067,581 train instances and 13,985 development instances (without reduplication)

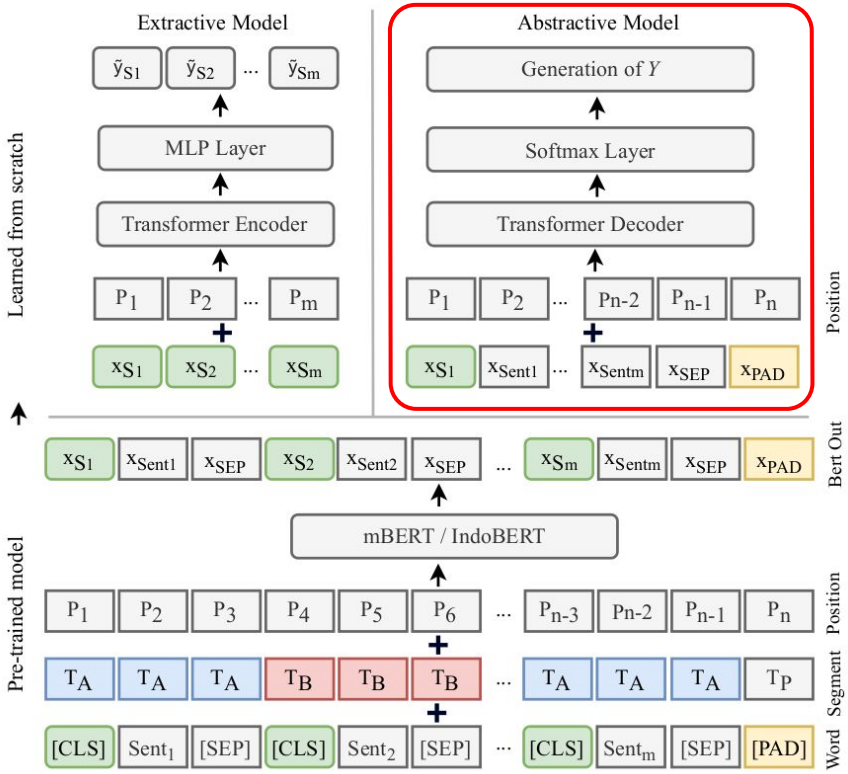
→ trained the model for 2.4M steps (180 epochs) for a total of 2 calendar months,
→ the final perplexity over the development set being **3.97 (similar to English BERT-base)**

mBERT and IndoBERT for Summarization Model



- Extractive label (ORACLE) is created by greedily optimize ROUGE-1.
- **BertExt** (extractive model)

mBERT and IndoBERT for Summarization Model

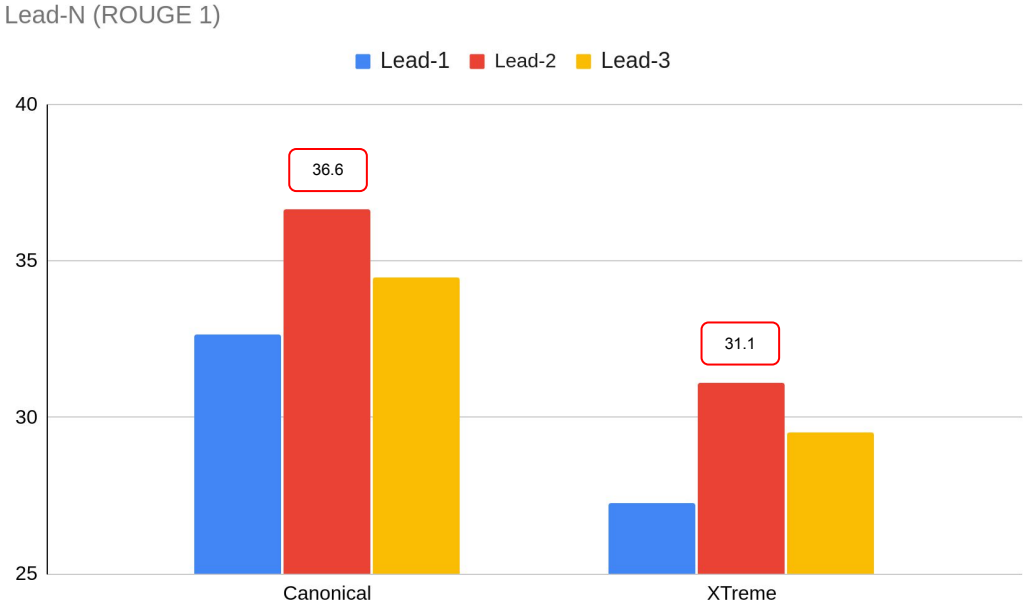


- **BertAbs** (abstractive model)
- **BertExtAbs** (extractive + abstractive model)

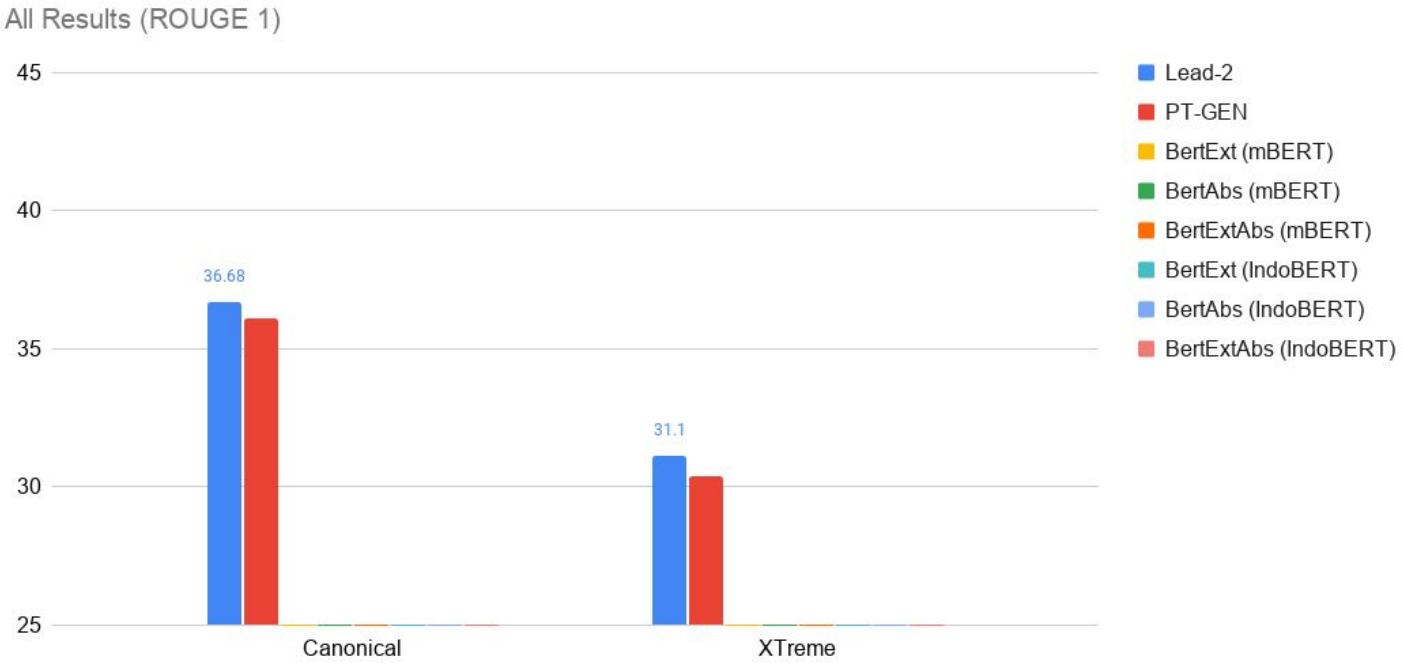
4

Experiment Results

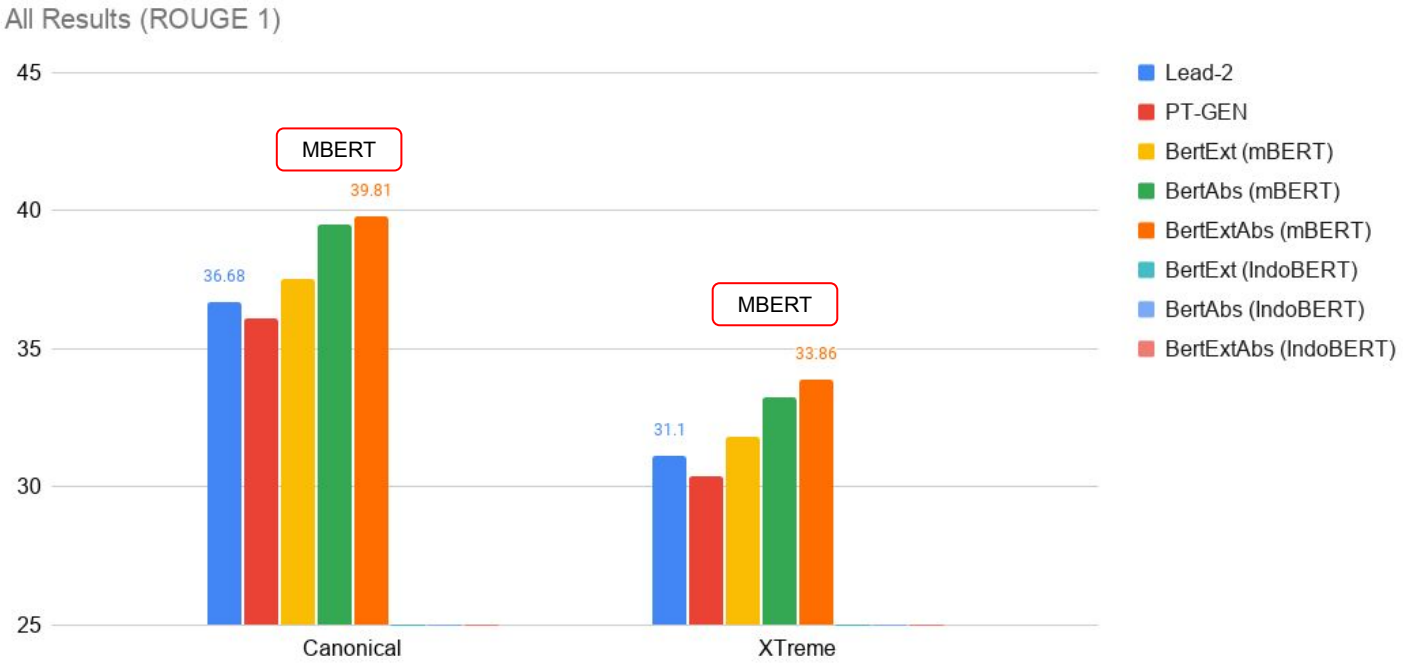
Lead-N


Lead-N (ROUGE 1)
Lead-1
Lead-2
Lead-3
40
35
30
25
36.6
31.1
Canonical
XTreme

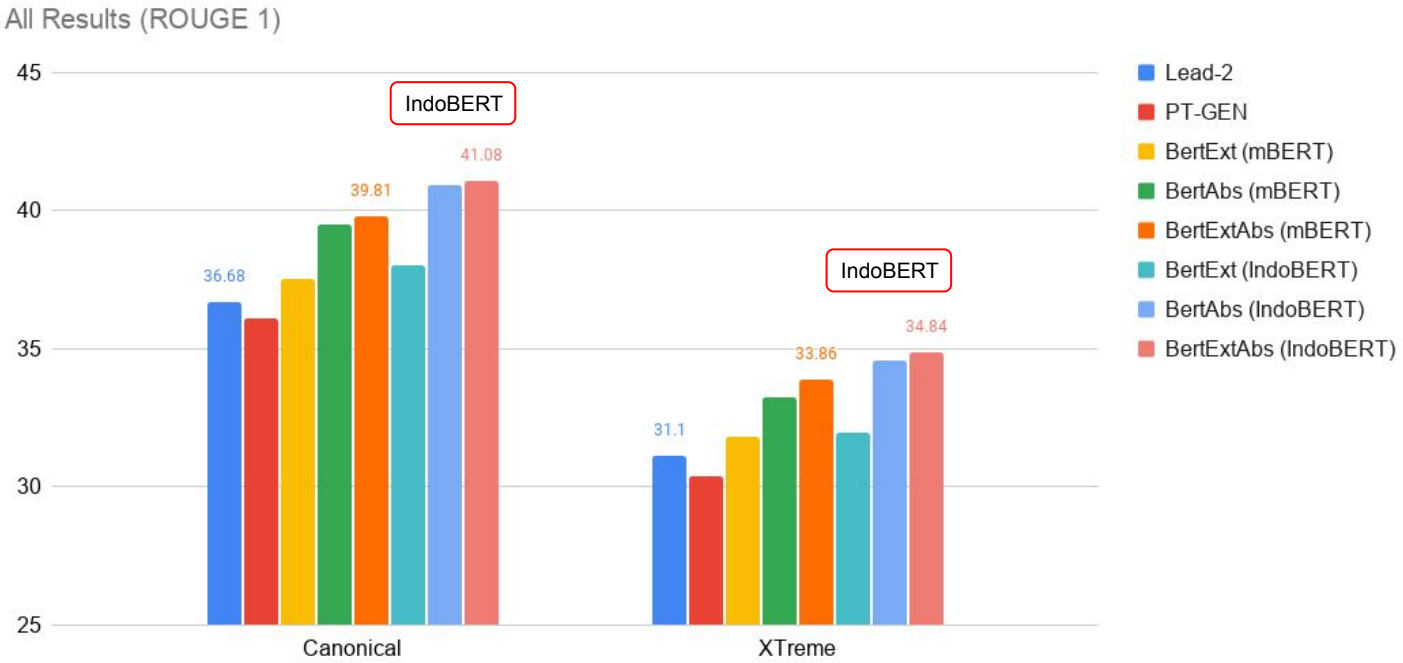
All Results


All Results (ROUGE 1)
45
40
35
30
25
36.68
31.1
Canonical
XTreme
Lead-2
PT-GEN
BertExt (mBERT)
BertAbs (mBERT)
BertExtAbs (mBERT)
BertExt (IndoBERT)
BertAbs (IndoBERT)
BertExtAbs (IndoBERT)

All Results


All Results (ROUGE 1)
45
40
35
30
25
MBERT
39.81
36.68
MBERT
33.86
31.1
Canonical
XTreme
Lead-2
PT-GEN
BertExt (mBERT)
BertAbs (mBERT)
BertExtAbs (mBERT)
BertExt (IndoBERT)
BertAbs (IndoBERT)
BertExtAbs (IndoBERT)

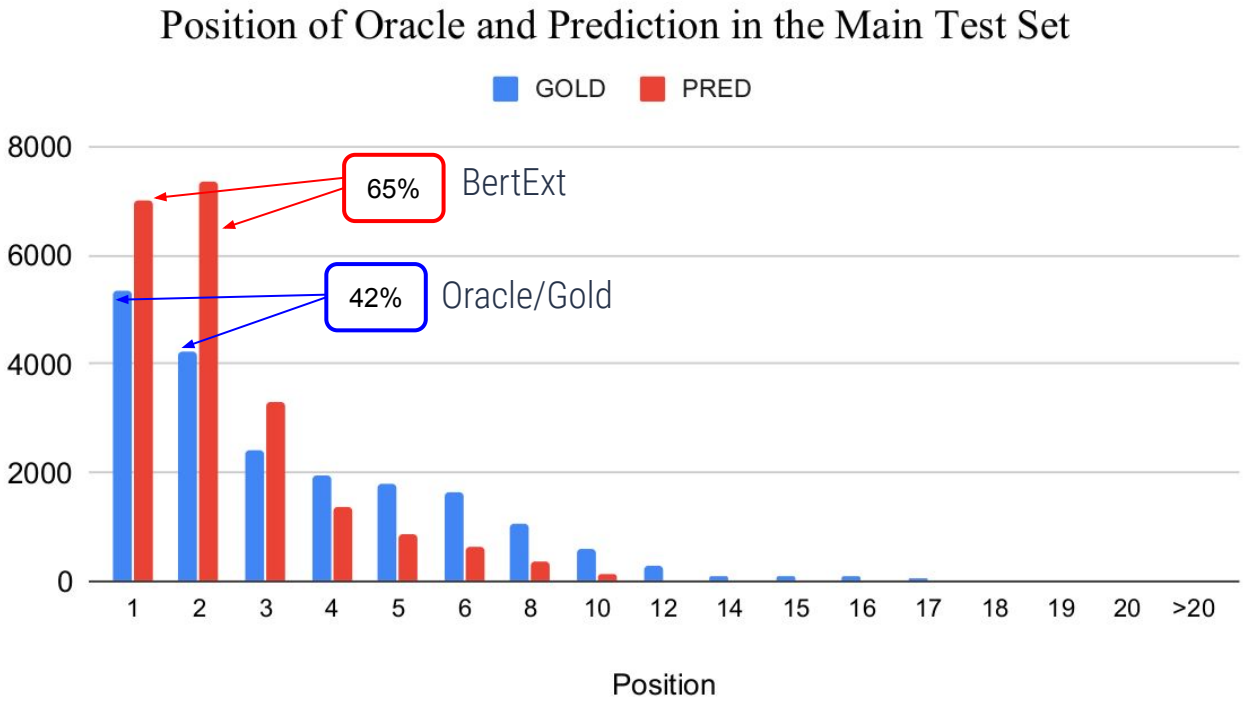
All Results


All Results (ROUGE 1)
45
40
35
30
25
IndoBERT
41.08
39.81
36.68
IndoBERT
34.84
33.86
31.1
Canonical
XTreme
Lead-2
PT-GEN
BertExt (mBERT)
BertAbs (mBERT)
BertExtAbs (mBERT)
BertExt (IndoBERT)
BertAbs (IndoBERT)
BertExtAbs (IndoBERT)

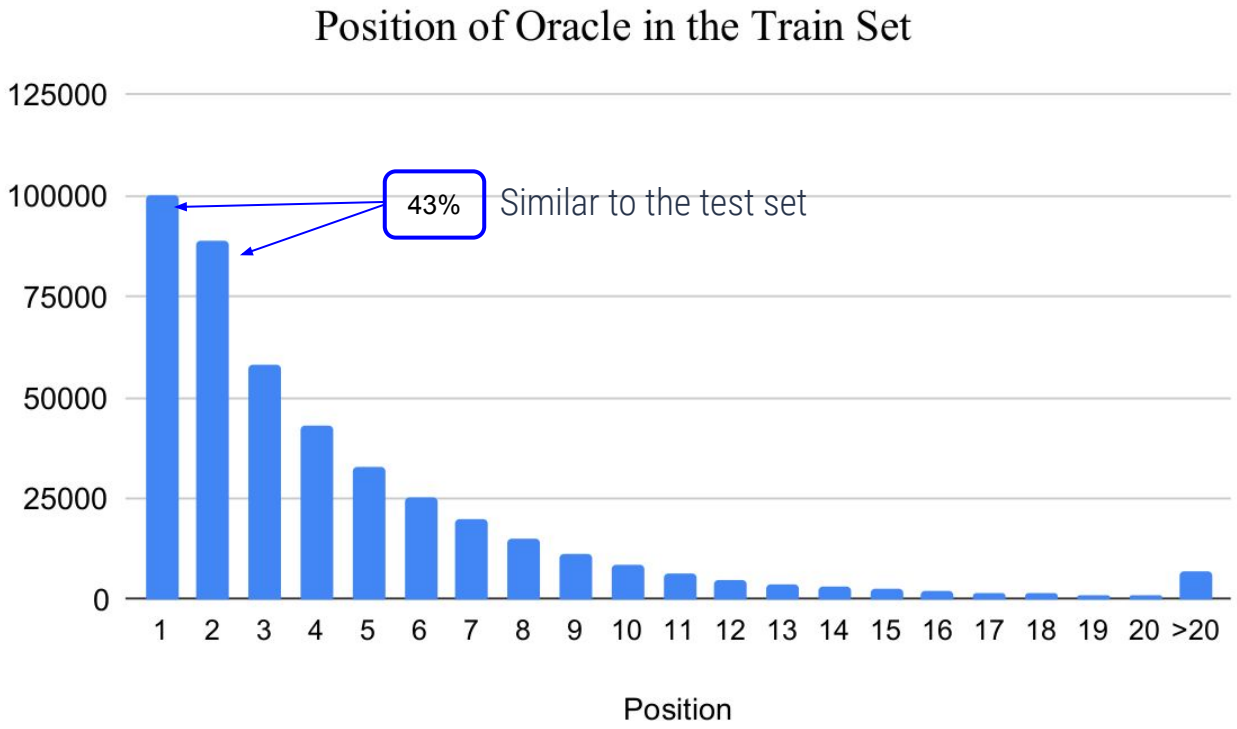
5

Error Analysis

A. Extractive Model (BertExt)



A. Extractive Model (BertExt)



B. Abstractive Model (BertExtAbs)

- We analyze test set with R1 score less than 0.4 (5,773 documents (nearly 50%)).

Procedure:

- Randomly select 100 samples.
- Two native Indonesian speakers annotate general quality: **1) bad, 2) average and 3) good.**
- Annotators read the article, gold summary, and the system summary

Pearson correlation among annotators: 0.692

B. Abstractive Model (BertExtAbs)

Fine-grained attributes for Error Analysis:

- Abbreviation → System summary uses abbreviations that are different to the reference
- Morphology → System summary uses morphological variants of the same lemmas
- Synonyms/paraphrasing
- Lack of Coverage
- Wrong Focus
- Unnecessary details (from document)
- Unnecessary details (not from document)

B. Abstractive Model (BertExtAbs)

Category	Bad	Avg.	Good
#Samples (100)	32	8	60
Abbreviation (%)	21.9	25.0	40.0
Morphology (%)	12.5	25.0	36.7
Paraphrasing (%)	50.0	87.5	**86.7**
Lack of coverage (%)	**90.6**	**100.0**	40.0
Wrong focus (%)	68.8	0.00	8.3
Un. details (from doc) (%)	**90.6**	75.0	75.0
Un. details (not from doc) (%)	18.8	12.5	5.0

B. Abstractive Model (BertExtAbs)

Category	Bad	Avg.	Good
#Samples (100)	32	8	60
Abbreviation (%)	21.9	25.0	40.0
Morphology (%)	12.5	25.0	36.7
Paraphrasing (%)	50.0	87.5	**86.7**
Lack of coverage (%)	**90.6**	**100.0**	40.0
Wrong focus (%)	68.8	0.00	8.3
Un. details (from doc) (%)	**90.6**	75.0	75.0
Un. details (not from doc) (%)	18.8	12.5	5.0

B. Abstractive Model (BertExtAbs)

Example-1 of error analysis (Abbreviation, morphoplogy, synonyms/paraphrashing, and details from the document)	
Dokumen: Liputan6.com , Jakarta : Protes masih bergema menyambut Keputusan Menteri Tenaga Kerja dan Transmigrasi Nomor 78 Tahun 2001 . Kebijakan yang sengaja dikeluarkan sebagai wujud perubahan keputusan sebelumnya ini , sampai sekarang , masih mengundang kecaman keras dari pekerja di Indonesia . Itulah sebabnya , mereka menuntut Kepmenakertrans baru ini dicabut karena dinilai merugikan pekerja . [19 kalimat dengan 406 kata tidak ditampilkan] Sementara itu , SPSI secara tegas menolak segala bentuk negosiasi . [3 kalimat dengan 45 kata setelahnya tidak ditampilkan] **Ringkasan manusia:** pemberlakuan kepmenakertrans 78/2001 masih mengundang rasa tidak puas di dada sejumlah pekerja indonesia . maka , lahirlah tuntutan agar peraturan yang dinilai merugikan ini dicabut . **Ringkasan sistem [Good]:** keputusan menteri tenaga kerja dan transmigrasi nomor 78 tahun 2001 mengundang kecaman keras dari pekerja di indonesia . mereka menuntut kepmenakertrans dicabut karena dinilai merugikan pekerja . spsi menolak negosiasi .	**Document:** Liputan6.com, Jakarta: Protests still resonate with welcoming Minister of Manpower and Transmigration Decree No. 78/2001. This policy, which was deliberately issued as an amendment to the previous decision, until now, still invites harsh criticism from workers in Indonesia. That is why they demand to revoke the new Kepmenakertrans because it is considered detrimental to workers. [19 sentences with 406 words are abbreviated from here] Meanwhile, SPSI firmly rejected all forms of negotiation. [3 sentences with 45 words are abbreviated from here] **Gold Summary:** The enactment of Kepmenakertrans 78/2001 still invites the dissatisfaction of Indonesian workers. hence, demands to revoke the regulation arose as it was considered to be detrimental. **System Summary [Good]:** Minister of Manpower and Transmigration Decree number 78 of 2001 invited strong criticism from workers in Indonesia. They demand to revoke Kepmenakertrans because it is considered detrimental to workers. SPSI rejects negotiations.

5

Conclusion

Key Takeaways

Our primary contributions:

(1) we release Liputan6, a large-scale Indonesian summarization corpus.
- ➔ train/val/test set is 193,883/10,792/10,792
- ➔ Xtreme val/test set is 4,948/3,862

(2) we create strong baseline with IndoBert
(3) we perform a thorough error analysis for future work.

THANKS!

https://github.com/fajri91/sum_liputan6