LipKey: A Large-Scale News Dataset for Absent Keyphrases Generation and Abstractive Summarization

Fajri Koto♠ Timothy Baldwin♠♥ Jey Han Lau♠
♠ School of Computing and Information Systems, The University of Melbourne
♥ Department of Natural Language Processing, MBZUAI
fajri.koto91@gmail.com tb@ldwin.net jeyhan.lau@gmail.com

Abstract

Summaries, keyphrases, and titles are different ways of concisely capturing the content of a document. While most previous work has released the datasets of keyphrases and summarization separately, in this work, we introduce LipKey, the largest news corpus with human-written abstractive summaries, absent keyphrases, and titles. We jointly use the three elements via multi-task training and training as joint structured inputs, in the context of document summarization. We find that including absent keyphrases and titles as additional context to the source document improves transformer-based summarization models.

1 Introduction

Key content of an article can be presented in different ways, including summaries, keyphrases, and titles. While most previous research has addressed each element individually (e.g. news summarization (Zhang et al., 2020a; Lewis et al., 2020; Koto et al., 2020a) and keyphrase generation in the scientific domain (Meng et al., 2017, 2021)), in this work we release a novel news dataset that consists of highly absent keyphrases (i.e. keyphrases which abstract over the content of the document), abstractive summaries, and titles to investigate the three elements in the context of single-document abstractive summarization.

Previous work has mainly utilized present keyphrases (i.e. keyphrases that are directly drawn from the source text) through unsupervised and supervised methods for summarization. For instance, traditional summarization models (Zhang et al., 2004; D’Avanzo and Magnini, 2005; Wan et al., 2007; Riedhammer et al., 2010; Qazvinian et al., 2010) and modern neural models (Müngen and Kaya, 2018; Nallapati et al., 2016; Liu et al., 2021) have been combined with the top-k frequent words, TF-IDF, and TextRank (Mihalcea and Tarau, 2004) to obtain keyphrases. Elsewhere, Gehrmann et al. (2018); Li et al. (2020) used words contained in both the summary and article as keyphrases to improve summarization.

This paper aims to study how absent keyphrases (i.e. keyphrases that do not match any words in the source text) can be incorporated into summarization systems. Compared to present keyphrases used in previous work, absent keyphrases potentially better complement abstractive summarization methods. Previous work has been hindered by the unavailability of a large annotated dataset with gold-standard summaries and keyphrases, thus opting for present keyphrase extraction (Qazvinian et al., 2010; Liu et al., 2021).

We additionally study the utility of titles in summarization. The underlying hypothesis is that titles and keyphrases are concise, complementary representations of an article, and provide relevant clues for summarization. While previous summarization datasets such as CNN/DM (Hermann et al., 2015), NYT (Sandhaus, 2008), and XSUM (Narayan et al., 2018) do not include keyphrases and titles, we present a novel large-scale dataset containing both.

Following Koto et al. (2020a), we crawl Liputan62—an Indonesian news portal — to obtain 105K news articles with titles, abstractive summaries, and absent keyphrases, all authored by journalists. Note that the dataset of Koto et al. (2020a) is based on the time period 2000–2010, at which point Liputan6 did not include keyphrases, while our dataset is based on the time period 2019–2021.3 Furthermore, the fact that the dataset is in Indonesian contributes to language diversity in NLP (Joshi et al., 2020).

2https://www.liputan6.com
3Koto et al. (2020a) also do not release the titles. We performed online crawling using an RSS feed taken from a two year period to obtain the dataset.
To summarize our contributions: (1) we release LipKey, the largest news corpus containing human-written abstractive summaries and absent keyphrases, as well as being the first large-scale Indonesian keyphrase dataset; (2) through extensive experimentation, we benchmark multi-task training and structured input methods using keyphrases and titles for Indonesian text summarization over different pretrained language models. We find that incorporating keyphrases and titles as structured inputs performs better than multi-task training, and consistently improves summary quality.

## 2 Related Work

Most keyphrase datasets are in the domain of English scientific publications (Hulth, 2003; Krapivin et al., 2009; Kim et al., 2010; Meng et al., 2021). In Table 1, we compare our corpus, LipKey, with other keyphrase datasets in the news domain. Most datasets such as DUC–2001 (Wan and Xiao, 2008), PT–BN–KP (Marujo et al., 2012), KPCrowd (Marujo et al., 2011), and WikiNews (Bougouin et al., 2013) are small in size and consist of highly present keyphrases, with KPTimes (Gallina et al., 2019) being the only exception. DUC–2001 is the only dataset with both keyphrases and summaries, but has only 308 documents. In comparison, LipKey is a large news corpus that includes human-written summaries and absent keyphrases, as well as being the first large-scale Indonesian keyphrase dataset.

Incorporating keyphrases into summarization has been explored in other languages such as Chinese (Jiang et al., 2018; Mihalcea and Tarau, 2004), but using present keyphrases. This is the first work to combine the two tasks in the Indonesian language, with previous work separately tackling: (1) keyphrase extraction, over Twitter (Mahfuzh et al., 2019), consumer-health questions (Saputra et al., 2018), or scientific articles (Asrori et al., 2020; Trisna and Nurwidayantoro, 2020) with limited data; or (2) document summarization in the news domain (Kurniawan and Louvan, 2018; Koto et al., 2020a).

## 3 Data Construction

Liputan6 is one of the largest Indonesian news portals, containing news on topics such as politics, health, business, and popular culture. Koto et al. (2020a) found that Liputan6 summaries are highly abstractive, written by journalists, and suitable for Indonesian text summarization research. The summary and keyphrases are encapsulated in javascript variables window.kmklabs.article with the keys shortDescription and keywords, respectively. In crawling Liputan6, we use article ID to ensure there is no redundancy in the dataset. LipKey articles span the period December 2019 to March 2021, and each article is associated with a summary, title, and keyphrase(s). In Table 2 and Table 3, we show the overall data statistics of LipKey, and compare it with previous Indonesian

---

**Table 1:** LipKey and other keyphrase datasets in the news domain. “AbsKey” is the percentage of “absent” keyphrases, relative to the source article.

| Dataset / Lang | Size | Includes Summ? | #Key per doc (%) | AbsKey (%) |
|----------------|------|----------------|------------------|------------|
| LipKey (ours) / id | 105,537 | Yes | 4.5 | 51.2 |
| DUC-2001 / en | 308 | Yes | 8.1 | 3.7 |
| PT-BN-KP / en | 110 | No | 23.7 | 2.5 |
| KPCrowd / en | 500 | No | 48.9 | 13.5 |
| KPTimes / en | 289,923 | No | 5.0 | 54.8 |
| WikiNews / fr | 100 | No | 11.8 | 5.0 |

**Table 2:** Per-article summary statistics for LipKey. For keyphrases, #sentence indicates #keyphrases.

| Dataset | Size | Article | Title | Summary | Keyphrases |
|---------|------|---------|-------|---------|------------|
| IndoSum | 18,764 | 3.1 | 10.8 | 16.2 | 20.3 |
| Liputan6 | 215,827 | 12.9 | 41.6 | 57.6 | 66.9 |
| LipKey (summary) | 105,537 | 7.5 | 25.2 | 35.1 | 40.9 |
| LipKey (title) | 105,537 | 26.8 | 65.4 | 84.5 | 92.7 |

**Table 3:** Abstractiveness of summaries (and titles) in IndoSum, Liputan6, and LipKey, compared to the article.

| Dataset | Size | % of novel n-gram |
|---------|------|-------------------|
| IndoSum | 18,764 | 1.3 | 8.1 | 16.2 | 20.3 |
| Liputan6 | 215,827 | 12.9 | 41.6 | 57.6 | 66.9 |
| LipKey (summary) | 105,537 | 7.5 | 25.2 | 35.1 | 40.9 |
| LipKey (title) | 105,537 | 26.8 | 65.4 | 84.5 | 92.7 |

---

None of the datasets are publicly available.

According to https://www.alexa.com, Liputan6 was ranked 16th and 308th in Indonesia and worldwide, respectively, in November 2021 in terms of popularity.

In 2012, Liputan6 added keyphrases for articles. These keyphrases are also assigned manually by the journalist.

Since the data is crawled between December 2019 and March 2021, models trained on this data will likely be biased towards events and issues occurring in this period.
summarization datasets: IndoSum (Kurniawan and Louvan, 2018) and Liputan6 (Koto et al., 2020a).

We observe that summaries in LipKey are more abstractive than IndoSum in terms of novel n-grams (computed relatively to the article). Interestingly, we found that LipKey’s titles are even more abstractive than the summaries in all datasets. Note that the median summary length in LipKey is one sentence, and shorter than Liputan6 (Koto et al., 2020a) at two sentences, despite both datasets being crawled from the same news portal.

In constructing LipKey, we discard instances where: (1) one of the keyphrases has more than 6 words (which tends to be noise); (2) the article has less than 15 words; or (3) the summary has less than 5 words. This results in 105,537 instances that we split into 96,541/4,154/4,842 for train/development/test. In terms of the number of words, 33% and 43% of keyphrases consist of 1 and 2 words, respectively, with the remainder being 3–6 words (see Table 10 in the Appendix for some examples).

We also perform manual analysis over 100 random samples to examine why the keyphrases are absent (i.e. do not occur) in the source article. We find that 80% of keyphrases partially match the article or are word-order variants (see Table 5). Moreover, 15%, 12% and 14% of absent keyphrases are acronyms, synonyms, or morphological variants.

LipKey consists of diverse news genres as shown in Table 4. “General”, “sport”, “business”, “local”, and “entertainment” are the top-5 most common news genres found in the dataset, covering 75% of all articles. We observe that “entertainment” articles tend to be shorter than other genres, and “automotive” has the longest summaries but the fewest keyphrases on average. The average edit distance between two pairs of keyphrases in each genre is almost similar, ranging between 10–13, indicating the diversity of keyphrases in each article. Lastly, word-level entropy in each genre is also similar (around 10) indicating the similar low-level redundancy in each news genre.

Table 4: Data statistics based on news genre. ED is the average character-level Levenshtein edit distance, computed between two pairs of keyphrases, while word-level entropy (1-gram) is calculated based on the concatenation of article, title, and summary.

| Genre    | Total (%) | Article | Title | Summary | Keyphrase | Word-level |
|----------|-----------|---------|-------|---------|-----------|------------|
|          |           | (Vocab) | (µ(#word)) | (Vocab) | (µ(#word)) | (Vocab) | (µ(#word)) | µ | ED | Entropy |
| general  | 33.5%     | 155,834 | 421.6 | 29,155 | 31,935 | 18.8 | 4.3 | 11.6 | 10.8 |
| sport    | 12.3%     | 66,303 | 340.2 | 11,359 | 13,178 | 18.9 | 5.2 | 12.1 | 10.4 |
| business | 12.2%     | 82,899 | 524.7 | 13,377 | 15,591 | 18.3 | 4.1 | 10.8 | 10.5 |
| local    | 9.5%      | 76,417 | 391.0 | 14,266 | 18,969 | 21.9 | 4.8 | 13.3 | 10.5 |
| entertainment | 8.6% | 62,677 | 270.2 | 12,265 | 13,178 | 18.9 | 5.2 | 12.1 | 10.4 |
| lifestyle | 6.3%      | 75,566 | 434.7 | 14,206 | 12,265 | 21.9 | 4.8 | 11.6 | 10.7 |
| international | 5.5% | 61,007 | 460.4 | 9,822 | 11,545 | 19.4 | 5.7 | 11.8 | 10.7 |
| health   | 4.7%      | 44,916 | 380.8 | 8,314 | 9,468 | 19.2 | 3.7 | 11.1 | 10.7 |
| technology | 3.2% | 39,338 | 434.7 | 11,456 | 13,177 | 18.9 | 4.1 | 11.6 | 10.7 |
| automotive | 2.7% | 35,870 | 369.8 | 5,855 | 8,503 | 24.9 | 3.7 | 12.4 | 10.5 |
| other    | 1.5%      | 37,512 | 603.7 | 4,131 | 5,278 | 18.0 | 4.7 | 12.4 | 10.3 |

4 Experiments

4.1 Set-Up

As described in Figure 1, we experiment in two settings: (1) multi-task training (title/keyphrases = output); and (2) training with structured input (title/keyphrases = input). For the first, we use summary s, title t, and keyphrase(s) k as the separate target texts, and perform multi-task training with article a as the source text (thus three tasks: summarization, keyphrase generation (KPG), and title generation). The total loss \( L \) for multi-task
training is defined as $\mathcal{L}_s + \mathcal{L}_t + \mathcal{L}_k$. For the second, the goal is to learn $P(s|t, k, a)$ that is realized by concatenating title $t$, keyphrases $k$, and article $a$ to form the source text, and use summary $s$ as the target text. To distinguish the four text types and structure the input, we introduce the special tokens of [SUMMARY], [TITLE], [KEYPHRASES], and [ARTICLE] for all pretrained language models. In the case of multiple keyphrases, we use <sep> as a separator. The maximum number of tokens for the article is 512, and for the summary, title, and keyphrases it is 100.

We use the huggingface transformers library (Wolf et al., 2020) for our experiments with three pretrained language models: IndoBERT\(^8\) (Koto et al., 2020b), mT5 (base)\(^9\) (Xue et al., 2021), and mBART (large)\(^10\) (Liu et al., 2020). For the monolingual IndoBERT, we follow Liu and Lapata (2019) in adding a raw transformer decoder (layers = 6, hidden size = 768, feed-forward = 2,048, and heads = 8) on top of IndoBERT, and train it on 4×V100 16GB GPUs for 20K steps. For the multilingual mT5 and mBART, we train them on 4×V100 32GB GPUs for 60 epochs (around 20K steps) with an initial learning rate of 1e-4 (Adam optimizer). We pick the best checkpoint based on ROUGE scores (Lin, 2004) on the development set (see the Appendix for more details of hyper-parameters).

Additionally, we train keyphrase generation (KPG) models (Seq2Seq) with the same architectures and configurations as the summarization models. We compare the generated keyphrases with: (a) human-written keyphrases; and (b) keyphrases from RAKE, an unsupervised language-independent keyphrase extraction method (Rose et al., 2010).

---

\(^8\)indolem/indobert-base-uncased  
\(^9\)google/mt5-base  
\(^10\)facebook/mbart-large-50

### Table 6: Summarization results on LipKey. “Foc” and “Cov” are Focus and Coverage, respectively, of FFCI. Entries in bold and underline refer to the best overall score and the best score for each model, respectively. “Oracle” is obtained by greedily selecting the subset of sentences in the article that maximizes the ROUGE score based on the reference summary.

| Model                     | R1   | R2   | RL  | Foc  | Cov  |
|---------------------------|------|------|-----|------|------|
| Oracle                    | 69.2 | 58.9 | 66.9| 76.4 | 87.2 |
| mBART (large) with 600M parameters |      |      |     |      |      |
| mT5 (base) with 580M parameters |      |      |     |      |      |
| multi-task training       |      |      |     |      |      |
| summary, keyphrase        | 43.1 | 31.3 | 40.5| 67.6 | 73.9 |
| summary, title            | 42.2 | 30.0 | 39.5| 67.3 | 73.4 |
| summary, keyphrase, title | 43.5 | 31.6 | 40.8| 67.8 | 74.1 |
| training with additional context |      |      |     |      |      |
| + keyphrase               | 43.3 | 31.2 | 40.4| 67.7 | 73.8 |
| + title                   | 43.3 | 30.8 | 40.4| 67.7 | 73.8 |
| + keyphrase + title       | 44.8 | 32.3 | 42.0| 68.8 | 74.6 |
| m5 (base) with 153M parameters |      |      |     |      |      |
| Oracle                    | 69.2 | 58.9 | 66.9| 76.4 | 87.2 |
| Lead-1                    | 36.6 | 26.1 | 34.1| 58.5 | 71.8 |
| multi-task training       |      |      |     |      |      |
| summary, keyphrase        | 44.7 | 33.2 | 42.1| 66.9 | 76.3 |
| summary, title            | 44.6 | 33.1 | 42.0| 66.6 | 76.4 |
| summary, keyphrase, title | 43.7 | 32.0 | 41.0| 66.1 | 76.0 |
| training with additional context |      |      |     |      |      |
| + keyphrase               | 46.4 | 34.8 | 43.8| 68.2 | 76.6 |
| + title                   | 45.4 | 33.8 | 42.9| 67.5 | 76.4 |
| + keyphrase + title       | **46.7** | **35.1** | **44.2** | **68.4** | **76.9** |

For evaluating the summarization models, we use F1 of ROUGE scores (R1, R2, and RL), and Focus and Coverage from the FFCI framework (Koto et al., 2022), computed based on Precision and Recall of BERTSCORE (Zhang et al., 2020b) using mBERT uncased.\(^{11}\) For evaluating KPG, we use macro-averaged $F_1 @ 5$, $F_1 @ O$, and $F_1 @ M$, following Meng et al. (2021), and additionally report R1, Focus, and Coverage. Detailed definitions of the metrics are provided in the Appendix.

\(^{11}\)For details of BERT layer selection, see Koto et al. (2021).
4.2 Results

In Table 6, we show the full experimental results on the test set. First, we observe that vanilla models (trained only using the article) substantially outperform Lead-1 for all models. We find that the vanilla model of mT5 performs better than IndoBERT and mBART, with an improvement of +3.4 and +2.1 R1, respectively.

Training with additional context as structured input consistently improves over multi-task training, with the best results generally being obtained with both keyphrases and title, and mT5 being the best model. When incorporating each element separately, keyphrases are generally better than titles, improving over the vanilla model, with IndoBERT (with multi-task training) being the notable exception. We also observe that mBART (large) and mT5 (base) are similar in parameter size (600M), but mT5 is substantially better. The FFCI framework shows that both models have similar Focus (= precision), but mT5 has higher Coverage (= recall).

Next, in Table 7, we present results for keyphrase generation on the LipKey test set, and observe that mBART (large) achieves the best performance across all metrics. Interestingly, RAKE performs very poorly, in part emphasizing the limitations of the extractive RAKE method (vs. the highly absent keyphrases in LipKey).

Lastly, to benchmark the effect of different keyphrases in summarization we perform an ablation study over the best summarization model, mT5, using keyphrases sourced through three different methods: (1) RAKE, (2) Seq2Seq, and (3) human-assigned. We use RAKE for this study because there is no suitable keyphrase dataset for training neural models to extract present keyphrases in Indonesian. As seen in Table 8, adding RAKE keyphrases hurts summarization results, but when using Seq2Seq keyphrases (generated by mBART), the performance consistently improves across all metrics, close to the performance of human-assigned keyphrases. Considering this finding, it would be interesting to explore the transferability of keyphrase generation models to other languages, to see if it can be reproduced.

5 Conclusion

In this paper, we release LipKey, the largest news corpus with human-written keyphrases, summaries and titles which is also the first large scale Indonesian keyphrase dataset. We experimented with incorporating keyphrases (and titles) into summarization training via multi-task training or as structured inputs, and found that the latter works better. In this preliminary results, we show that absent keyphrases benefit summarization systems more than present keyphrases extracted by RAKE.

6 Ethical Considerations

According to Indonesian Copyright Law number 28 year 2014 article 44, the use, retrieval, reproduction, and/or change of works and/or related rights products in whole or substantial part are not regarded as a copyright infringement if the source is mentioned or cited in full for the purpose of education and research.

Acknowledgements

We are grateful to the anonymous reviewers for their helpful feedback and suggestions. The first author is supported by the Australia Awards Scholarship (AAS), funded by the Department of Foreign Affairs and Trade (DFAT), Australia. This research was undertaken using the LIEF HPC-GPGPU Facility hosted at The University of Melbourne. This

Table 7: Keyphrase generation results on LipKey.

| Model        | R1 | Foc. | Cov. | $F_1 \@ 5$ | $F_1 \@ \infty$ | $F_1 \@ M$ |
|--------------|----|------|------|------------|----------------|------------|
| RAKE         | 7.8| 40.0 | 58.7 | 1.0        | 1.0             | 1.0        |
| IndoBERT     | 58.8| 74.2 | 79.4 | 45.5       | 45.2            | 46.5       |
| mT5 (base)   | 62.0| 75.7 | 81.7 | 53.3       | 52.9            | 54.5       |
| mBART (large)| 63.4| 76.4 | 81.9 | 54.5       | 54.4            | 56.0       |

Table 8: Ablation study of mT5 (base) over different keyphrases on test set. * denotes using mBART (large).

| Model                        | R1  | R2  | RL  | Foc. | Cov. |
|------------------------------|-----|-----|-----|------|------|
| Vanilla                      | 45.2| 33.7| 42.7| 67.5 | 76.2 |
| + keyphrases (RAKE)          | 44.8| 33.3| 42.3| 66.5 | 75.6 |
| + keyphrases (Seq2Seq*)      | 46.0| 34.4| 43.5| 68.1 | 76.4 |
| + keyphrases (Human)         | 46.4| 34.8| 43.8| 68.2 | 76.6 |
| Vanilla + title              | 45.4| 33.8| 42.9| 67.5 | 76.4 |
| + keyphrases (RAKE)          | 43.7| 32.1| 41.2| 67.3 | 76.1 |
| + keyphrases (Seq2Seq*)      | 45.9| 34.1| 43.3| 67.9 | 76.4 |
| + keyphrases (Human)         | 46.7| 35.1| 44.2| 68.4 | 76.9 |

12We choose Lead-1 because the average #sentence of the summary is 1.2 in Table 2.

13For each article, we pick the top-5 keyphrases based on RAKE scoring.

14https://wipolex-res.wipo.int/edocs/lexdocs/laws/en/id/id064en.pdf
facility was established with the assistance of LIEF Grant LE170100200.

References

Riris Bayu Asrori, Robert Setyawan, and Muljono Muljono. 2020. Performance analysis graph-based keyphrase extraction in Indonesia scientific paper. In 2020 International Seminar on Application for Technology of Information and Communication (iSemantic).

Adrien Bougouin, Florian Boudin, and Béatrice Daille. 2013. TopicRank: Graph-based topic ranking for keyphrase extraction. In Proceedings of the Sixth International Joint Conference on Natural Language Processing, pages 543–551, Nagoya, Japan. Asian Federation of Natural Language Processing.

Ernesto D’Avanzo and Bernado Magnini. 2005. A keyphrase-based approach to summarization: the Lake system at DUC-2005. In Proceedings of DUC.

Ygor Gallina, Florian Boudin, and Beatrice Daille. 2019. KPTimes: A large-scale dataset for keyphrase generation on news documents. In Proceedings of the 12th International Conference on Natural Language Generation, pages 130–135, Tokyo, Japan. Association for Computational Linguistics.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4098–4109, Brussels, Belgium. Association for Computational Linguistics.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In NeurIPS’15 Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1, volume 28, pages 1693–1701.

Anette Hulth. 2003. Improved automatic keyword extraction given more linguistic knowledge. In Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, pages 216–223.

Xiaoping Jiang, Po Hu, Liwei Hou, and Xia Wang. 2018. Improving pointer-generator network with keywords information for Chinese abstractive summarization. In CCF International Conference on Natural Language Processing and Chinese Computing, pages 464–474.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6282–6293, Online. Association for Computational Linguistics.

Su Nam Kim, Olena Medelyan, Min-Yen Kan, and Timothy Baldwin. 2010. Semeval-2010 task 5 : Automatic keyphrase extraction from scientific articles. In Proceedings of the 5th International Workshop on Semantic Evaluation, pages 21–26.

Fajri Koto, Timothy Baldwin, and Jey Han Lau. 2022. FFCI: A framework for interpretable automatic evaluation of summarization. Journal of Artificial Intelligence Research, 73:1553–1607.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2020a. Liputan6: A large-scale Indonesian dataset for text summarization. In Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, pages 598–608, Suzhou, China. Association for Computational Linguistics.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. Evaluating the efficacy of summarization evaluation across languages. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 801–812, Online. Association for Computational Linguistics.

Fajri Koto, Afshin Rahimi, Jey Han Lau, and Timothy Baldwin. 2020b. IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. In Proceedings of the 28th International Conference on Computational Linguistics, pages 757–770, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Mikalai Krapivin, Aliaksandr Autaeu, and Maurizio Marchese. 2009. Large dataset for keyphrases extraction. Technical Report DISI-09-055, University of Trento.

Kemal Kurniawan and Samuel Louvan. 2018. Indosum: A new benchmark dataset for Indonesian text summarization. In Proceedings of the 2018 International Conference on Asian Language Processing (IALP), pages 215–220.

Mike Lewis, Yinhua Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7871–7880, Online. Association for Computational Linguistics.

Haoran Li, Junning Zhu, Jiajun Zhang, Chengqiu Zong, and Xiaodong He. 2020. Keywords-guided abstractive sentence summarization. In Proceedings of the AAAI Conference on Artificial Intelligence, pages 8196–8203.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In Text Summarization Branches Out, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Yinhao Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. Transactions of the Association for Computational Linguistics, 8:726–742.

Yizhu Liu, Qi Jia, and Kenny Zhu. 2021. Keyword-aware abstractive summarization by extracting set-level intermediate summaries. In Proceedings of the Web Conference 2021, pages 3042–3054.

Miftahul Mahfuzh, Sidik Soleman, and Ayu Purwarianti. 2019. Improving joint layer RNN based keyphrase extraction by using syntactical features. In 2019 International Conference of Advanced Informatics: Concepts, Theory and Applications (ICAICTA).

Luís Marujo, Anatoile Gershman, Jaime Carbonell, Robert Frederking, and João P. Neto. 2012. Supervised topical key phrase extraction of news stories using crowdsourcing, light filtering and co-reference normalization. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12), pages 399–403, Istanbul, Turkey. European Language Resources Association (ELRA).

Luís Marujo, Márcio Viveiros, and João Paulo da Silva Neto. 2011. Keyphrase cloud generation of broadcast news. In Interspeech 2011, pages 2393–2396.

Rui Meng, Xingdi Yuan, Tong Wang, Sanqiang Zhao, Adam Trischler, and Daqing He. 2021. An empirical study on neural keyphrase generation. In Proceedings of the 2021 Conference of the North American Chapter for the Association for Computational Linguistics: Human Language Technologies, pages 4985–5007, Online. Association for Computational Linguistics.

Rui Meng, Sanqiang Zhao, Shuqiang Han, Daqing He, Peter Brusilovsky, and Yu Chi. 2017. Deep keyphrase generation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 582–592, Vancouver, Canada. Association for Computational Linguistics.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Ahmet Anil Müngen and Mehmet Kaya. 2018. Extracting abstract and keywords from context for academic articles. Social Network Analysis and Mining, 8(1):45.

Ramesh Nallapati, Bowen Zhou, Cicero Nogueira dos santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning, pages 280–290.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don’t give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Vahed Qazvinian, Dragomir R. Radev, and Arzu- can Özgür. 2010. Citation summarization through keyphrase extraction. In Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 895–903, Beijing, China. Coling 2010 Organizing Committee.

Korbinian Riedhammer, Benoit Favre, and Dilek Hakkani-Tür. 2010. Long story short — global unsupervised models for keyphrase based meeting summarization. Speech Communication, 52(10):801–815.

Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. Automatic keyword extraction from individual documents. Text Mining: Applications and Theory, pages 1–20.

Evan Sandhaus. 2008. The New York Times annotated corpus. Linguistic Data Consortium, Philadelphia, 6(12):e26752.

Ilham Fathy Saputra, Rahmad Mahendra, and Alfan Farizki Wicaksono. 2018. Keyphrases extraction from user-generated contents in healthcare domain using long short-term memory networks. In Proceedings of the BioNLP 2018 workshop, pages 28–34, Melbourne, Australia. Association for Computational Linguistics.

Nyoman Prayana Trisna and Arif Nurwidiantoro. 2020. Single document keywords extraction in Bahasa Indonesia using phrase chunking. TELKOMNIKA Telecommunication Computing Electronics and Control, 18(4):1917–1925.

Xiaojun Wan and Jianguo Xiao. 2008. Single document keyphrase extraction using neighborhood knowledge. In AAAI’08 Proceedings of the 23rd national conference on Artificial intelligence - Volume 2, pages 855–860.

Xiaojun Wan, Jianwu Yang, and Jianguo Xiao. 2007. Towards an iterative reinforcement approach for simultaneous document summarization and keyword
Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics, pages 483–498, Online. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020a. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In ICML 2020: 37th International Conference on Machine Learning, volume 1, pages 11328–11339.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. BERTScore: Evaluating text generation with BERT. In ICLR 2020: Eighth International Conference on Learning Representations.

Yongzheng Zhang, Evangelos Milios, and Nur Zincir-Heywood. 2004. A comparison of keyword-and keyterm-based methods for automatic web site summarization. In AAAI04 Workshop on Adaptive Text Extraction and Mining, pages 15–20.

### A Overview and Analysis of Keyphrases

| Text    | % Avg. of present keyphrases |
|---------|-------------------------------|
|         | Keyphrases level | Word level |
| Article | 66.8               | 86.9       |
| Summary | 39.8               | 55.3       |
| Title   | 48.2               | 62.9       |

Table 9: Proportion of keyphrases which match article, summary, and title

### B Training configurations

Summarization and keyphrase generation use the same models and architecture. For IndoBERT, we follow the Liu and Lapata (2019) architecture by adding a raw transformer decoder (layers = 6, hidden size = 768, feed-forward = 2,048, and heads = 8) on top of IndoBERT, and train it on 4 × V100 16GB GPUs for 200K steps with the Adam optimizer and learning rate \( lr = 2 \times 10^{-3} \cdot \min(\text{step} - 0.5, \text{step} \cdot 20, 000, 1.5) \) and 0.1 \( \cdot \min(\text{step} - 0.5, \text{step} - 10, 000, 1.5) \) for IndoBERT and the transformer decoder, respectively. We use a warmup of 20,000, a dropout of 0.2, a batch size total of 200 (10 x 4 GPUs x gradient accumulation of 5), and save checkpoints every 10,000 steps. We compute ROUGE scores (R1) to pick the best checkpoint based on the development set.

For mT5 and mBART, we train them on 4 × V100 32GB GPUs for 60 epochs (around 20K steps) with an initial learning rate of 1e-4 (Adam optimizer). We use a total batch size of 400 (10 x 4 GPUs x gradient accumulation of 10), a warmup of 10% of total steps, and save checkpoints every 1,000 steps. We also compute ROUGE scores (R1) to pick the best checkpoint based on the development set.

### C Evaluation Metrics

For summarization, we use ROUGE scores (Lin, 2004), and Focus and Coverage from the FFCI framework (Koto et al., 2022). Following Koto et al. (2021), for non-English text, focus and coverage are computed based on the precision and recall of BERTSCORE (Zhang et al., 2020b) using mBERT uncased at layers 12 and 6, respectively. For \( Y \) and \( Y' \) as the reference and system summary,
BERTScore is computed as follows:

\[
\mathcal{P}_{BERT} = \frac{1}{|Y|} \sum_{s_j \in Y} \max_{t_i \in Y'} t_i^T s_j
\]

\[
\mathcal{R}_{BERT} = \frac{1}{|Y|} \sum_{t_i \in Y'} \max_{s_j \in Y} t_i^T s_j
\]

\[
\mathcal{F}_{BERT} = 2 \frac{\mathcal{P}_{BERT} \cdot \mathcal{R}_{BERT}}{\mathcal{P}_{BERT} + \mathcal{R}_{BERT}}
\]

where \(s_j\) and \(t_i\) are token embeddings of \(Y\) and \(Y'\).

For evaluating the keyphrase generation model, we use macro-averaged \(F_1@5\), \(F_1@O\), and \(F_1@M\), following Meng et al. (2021). Given gold-standard keyphrases \(Y\) and the prediction \(\hat{Y} = \{y'_1, ..., y'_m\}\), we truncate the prediction to \(\hat{Y} = \{y'_1, ..., y'_{\min(k,m)}\}\) when only the top \(k\) predictions are used for evaluation. Precision, Recall, and \(F_1\) are consequently conditioned on \(k\), and computed as follows:

\[
P@k = \frac{|\hat{Y}_k \cap Y|}{|\hat{Y}_k|}
\]

\[
R@k = \frac{|\hat{Y}_k \cap Y|}{|Y|}
\]

\[
F_1@k = \frac{2 \cdot P@k \cdot R@k}{P@k + R@k}
\]

Thus \(F_1@5\) is \(F_1@k\) when \(k = 5\), \(F_1@O\) is \(F_1@k\) when \(k\) is the number of oracle (ground truth) keyphrases, and \(F_1@M\) is when \(k = |\hat{Y}|\).
The Ministry of Transportation (Kemenhub) ensures that the implementation of the protocol on the Jabodetabek Electric Circuit Train (KRL) continues. This statement was issued after 3 passengers from Bogor were tested positive for corona after a swab test was carried out. Spokesman for the Ministry of Transportation, Adita Irawati, stated that her party had issued Permenub No. 18/2020 which had regulated the operation of transportation modes during the pandemic. This is particularly the case in areas that have implemented Large-Scale Social Restrictions (PSBB) such as in Jabodetabek. It should be understood that the transmission of Covid-19 can occur anywhere, not only in KRL," said Adita, Tuesday (5/5/2020). First, passengers are required to wear masks. Second, he continued, officers check passengers’ body temperatures. [254 words are abbreviated from here]

Gold summaries:
The Ministry of Transportation stated that Permenhub 18/2020 has explicitly stated that there are several conditions that must be met by passengers of public transportation modes such as KRL.

IndoBERT:
The Ministry of Transportation (Kemenhub) ensures that the implementation of the protocol on the Jabodetabek Electric Circuit Train (KRL) continues.

mBART:
Permenhub 18/2020 has explicitly stated that there are several conditions that must be met by passengers of public transportation modes such as KRL.

mBART with additional contexts (+ keyphrases + titles):
Adita said that Permenhub 18/2020 has explicitly stated that there are several conditions that must be met by passengers of public transportation modes such as KRL.

mT5:
The Ministry of Transportation ensures that the implementation of the protocol on the Jabodetabek Electric Circuit Train (KRL) continues.

mT5 with additional contexts (+ keyphrases + titles):
The Ministry of Transportation ensures that the implementation of the protocol on the Jabodetabek Electric Circuit Train (KRL) continues.

Figure 2: Example from the LipKey dataset, with gold-standard and generated summaries.
| Indonesian | English (translation) |
|------------|----------------------|
| **Gold Keyphrases:** | **Gold Keyphrases:** |
| Relawan Uji Vaksin, Vaksin Novacov | Vaccine Test Volunteers, Sinovac Vaccines |
| **Article:** liputan6 , com , jakarta - manajer lapangan tim riset uji klinis vaksin covid-19 sinovac , dr eddy fadliyana menyebut sejauh ini sudah ada sekitar 1. 020 calon relawan yang mendaftarkan diri untuk mengikuti uji vaksin dari tlongkok . itu . dia mengatakan , pelaksanaan uji vaksin itu akan dilakukan selasa 11 agustus 2020 . pada hari pertama itu , uji vaksin bakal dilakukan di rumah sakit pendidikan ( rsp ) universitas padjadjaran , jalan eyckman , kota bandung ." sebutnya sama saja , hanya pemeriksaan di rsp itu , tes usapnya ( swab test ) didahului . sama saja sih prosedurnya , tidak ada yang berbeda , besok rsp imunisasi , kalau di tempat lain baru tahap awal ," kata eddy di bandung , senin ( 10/8/2020 ) . dikutip dari antara , menurut eddy , semua tempat yang ditunjuk menjadi lokasi uji vaksin covid-19 ini dipastikan sudah siap . mulai dari saran prasaranaan , menurutnya sudah sesuai dengan protokol kesehatan yang berlaku . dia mengatakan , uji vaksin itu dilakukan di enam lokasi , di antaranya yakni rps unpad , balai kesehatan unpad , puskesmas dago , puskesmas sukapakar , puskesmas garuda , dan puskesmas ciumbuleuit . dari seluruh calon relawan yang sudah mendaftar , menurutnya tak menutup kemungkinan sudah ada asy yang ikut mendaftar , karena , pendatangan untuk menjadi relawan itu terbuka untuk umum ." dari asy mungkin ada , saya tidak melihat statusnya apa pokoknya masyarakat yang mau silakan saja ," katanya . meski terbuka untuk umum , menurutnya ada beberapa syarat yang perlu dipenuhi oleh calon relawan antara lain usia relawan dalam rentang 18 hingga 59 tahun , dan dalam keadaan sehat tanpa penyakit bawaan . |
| **Article:** liputan6 , com , mamuju tengah - kejadian nasah menimpa h ( 40 ) warga desa barakkang , kecamatan budong-budong , mamuju tengah , sulawesi barat . ibu rumah tangga itu diterkam buaya saat mandi dan buang air besar di sungai . kapolsek budong-budong akp suparman menambahkan pristwa nasah itu ia mengatakan , peristiwa terjadi ini pada selasa ( 4/8/2020 ) dini hari , sekitar pukul 05 . 30 wita . korban yang tengah buang air besar itu tiba-tiba diterkam buaya yang memiliki panjang kurang lebih 7 meter , menurut saksi andi ( 38 ) yang merupakan adik korban , buaya itu tiba-tiba menerkam korban dari belakang . " kata suparman kepada liputan6 . com . petani labuhan batu utara diterkam buaya di depan anak istrinya suparman menambahkan , saksi juga sempat mendengarkan teriakan korban dan berusaha untuk menolong . namun , belum sempat menolong , buaya tersebut sudah terlebih dahulu menarik korban ke dalam air ." beberapa saat kemudian korban dan buaya muncul di permukaan air namun hanya sesaat lalu kemudian tenggelam lagi ke dalam air ." jelas suparman , hingga saat ini korban belum juga ditemukan . warga bersama pihak kepolisian sempat melakukan pencarian dengan peralatan seadanya . pihak bpbd mamuju tengah dan basarnas mamuju pun sudah dibungkahi ." saat ini bpbd dan masyarakat serta basasnas sudah ada di tpk melakukan pencarian ." tutup suparman . |

Figure 3: Example of articles and keyphrases in the LipKey dataset. We highlight words in the article that match its absent keyphrases with different colours. Yellow means partial match, green means acronym, and blue means morphological variants. The English translation is for illustration purposes.