



# Comparative Analysis of Retriever and Reader for Answering Open Domain Questions on BPS Knowledge in Indonesian

S P Widodo<sup>1,\*</sup>

<sup>1</sup> BPS-Statistics Indonesia, Jakarta, Indonesia

\*Corresponding author's email: sulisetyo.widodo@gmail.com

**Abstract.** Enumerators from Badan Pusat Statistik (BPS) still often encounter problems in finding solutions to cases encountered during censuses or surveys. Even though knowledge lists have been created and collected in various systems such as QA and knowledge management systems, enumerators still need to find appropriate answers from long and complex knowledge search results. On the other hand, Open-domain Question Answering (OpenQA) is capable of identifying answers to natural questions based on large-scale documents. OpenQA has main components, namely Retriever and Reader. For Retriever tasks, Dense Retrieval (DR) is proven to outperform traditional sparse retrieval such as TF-IDF or BM25. However, other research actually shows that BM25 is superior to DR in terms of accuracy. In this study, we compared DR and BM25 separately and DR+BM25 as a retriever. Additionally, we combine and evaluate several enhanced language models as Readers. In this way, a model with the best combination of Retriever and Reader can be obtained to be implemented in search systems such as QA and knowledge management systems.

## 1. Introduction

BPS provides data needs for government and public where data are obtained from censuses or surveys involving census officials. Every census official is required to obtain knowledge in form of concepts and definitions as well as case examples and solutions. However, quite a lot of problems were found in implementation of census. This is influenced by characteristics of an area and census itself. Efforts to document science have been made but there are still obstacles, where it is difficult to find right answers to questions from long and complex knowledge.

In general, Question Answering system (QA) is approach that is most likely to be able to overcome this problem. Where QA is a system that uses natural language questions to define user needs more specifically and naturally [1]. QA sub-section, namely Open-domain Question Answering (OpenQA), has ability to answer questions based on knowledge such as Freebase [2] and factual texts such as Wikipedia [3]. Currently, building OpenQA with the "Retriever-Reader" architecture has been recognized as the most efficient way [4]. Retrievers are tasked to retrieve documents relevant to given question, which can be considered as an IR (Information Retrieval) system, whereas Reader is tasked with inferring final answer from received document.

In Retriever function, Dense Passage Retrieval (DPR) can outperform powerful LuceneBM25 system by 9% - 19% in terms of top 20 passage retrieval accuracy [5]. DPR uses two independent BERT encoders (base, uncased) such as Open-Retrieval Question Answering (ORQA). However, DPR does



not require an expensive pre-training stage but instead focuses on learning a strong retriever using paired questions and answers. Meanwhile in another study combination of BERT Large Uncased with ElasticSearch (based on BM25) Retriever had a Correctly Answered score of 91.4% [6]. This score outperforms score of combination of uncased miniLM and DPR (Dence retrieval based) which only reached 85.2% for Correctly Answered. From these two studies a question arises. Is it possible to combine Dence retrieval and Bm25 as a Retriever task? So it will be interesting to see their ability to work together in OpenQA.

On other hand, an open source framework for OpenQA is available, namely haystack. In haystack, Dence and Bm25 Retriever methods are available. In addition, implementation of pretrained language model can be applied as a Reader on Haystack. This research explores retriever-reader architecture in case of getting answers to BPS knowledge in Indonesian. To get the best retriever-reader, several retrieval methods are compared and evaluated to get the best retriever. Then the best Retriever will be paired with several Reader models to get the best reader. Evaluation process uses an Indonesian language dataset obtained from Knowledge available at BPS. dataset is created following SQuAD format.

## 2. Literature Review

Information needs are often expressed as a question rather than a series of keywords. This condition is better known as Question Answering System (QA System). An Information retrieval system that allows users to ask questions naturally [1]. OpenQA is a sub-field of QA that can answer factoid questions by extracting knowledge from large collections of documents on diverse topics [6] [7] [8].

OpenQA has a sequential process flow, namely retrieving relevant documents, extracting candidate answers from retrieved documents, and reranking candidate answers to identify correct answer [9]. Modern OpenQA has an architecture known as "Retriever-Reader" [4] [10]. Retriever acts as an IR system whose purpose is to retrieve documents or related sections that may contain correct answer. Document retrieval is based on natural language queries which are then sorted according to their relevance. In general, there are three categories of Retrievers, namely Sparse Retrievers, Dense Retrievers, and Iterative Retrievers.

Sparse Retriever adopts classic IRs such as TF-IDF and BM25 as a method for searching relevant documents. drawback of TF-IDF and BM25 is that it is rare to measure match terms for document searches. In fact user questions often have terms that are not same as those that appear in document. Meanwhile, Dense Retriever, by adopting deep learning, can encode questions and documents into latent vector space. So that semantics of text outside of term match can be measured.

Furthermore, Karpukhin et al. [5] states that proposed Dence Passage Retriever (DPR) can outperform LuceneBM25 system by 9%-19% in terms of top 20 retrieval accuracy. DPR uses two independent encoders such as BERT to encode their respective questions and documents, and estimates their relevance by calculating a single similarity score between two representations. Using paired questions and answers DPR focused on designing a robust Retriever. So it does not require an expensive pre-training stage.

However, in another study, pairing of BERT Large Uncased with ElasticSearch(BM25) Retrievers had a Correctly Answered score of 91.4% [6]. This figure outperformed uncased miniLM and DPR pairs which only reached 85.2% for Correctly Answered. Other evaluation results also show that reducing top-k parameter can increase loss of answers and lead to poor performance of Question-Answering model. Although decreasing top-k Retrievers and top-k Readers can increase overall computation time.

Reader is main feature of modern OpenQA systems that differentiates QA systems from IR systems [4]. This is because Reader's job is to infer answers in response to questions from document being sorted. Reader receives input from document search results by Retriever. So Readers don't need to search for answers in complete document. An illustration of Retriever-Reader architecture in OpenQA is shown in Figure 1.

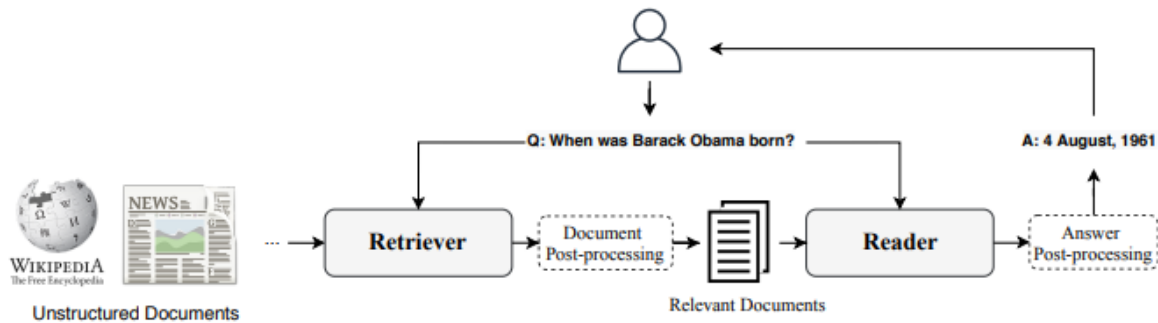


Figure 1. Retriever-Reader

### 3. Research Method

#### 3.1. Data Preparation

The dataset was obtained from knowledge available at BPS in form of Indonesian language training manuals. Knowledge is still available in files in formats such as PDF, Words, and Excel. So it is necessary to annotate knowledge and form a dataset in SQuAD format.

Annotations aim to define questions and answers that may appear in a context or knowledge. annotation process was carried out by three BPS employees as annotators using Haystack Annotation Tool. Figure 2. is an example of annotation process carried out by annotators on Haystack Annotation Tool. Meanwhile, Figure 3 shows results of annotation in form of a Json file in SQuAD format. annotated data consists of 95 contexts and is then used as a test dataset.

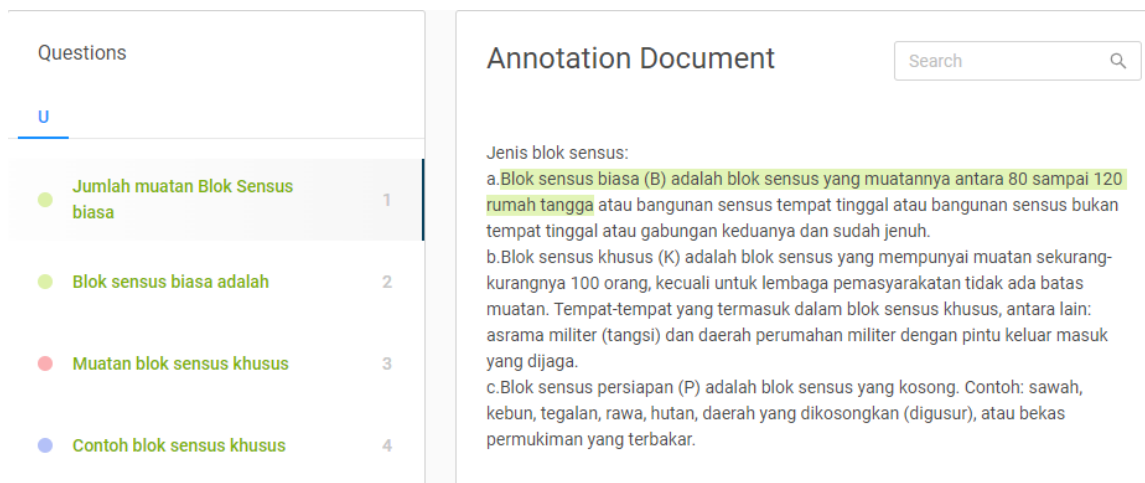


Figure 2. Knowledge annotation process in Haystack Annotation Tool



```

{
  "data": [
    {
      "paragraphs": [
        {
          "qas": [
            {
              "question": "Pengertian dari blok sensus adalah?",
              "id": 207613,
              "answers": [
                {
                  "answer_id": 262175,
                  "document_id": 393943,
                  "question_id": 207613,
                  "text": "Blok Sensus merupakan Bagian dari suatu wilayah desa/kelurahan yang merupakan daerah kerja dari seorang pencacah",
                  "answer_start": 0,
                  "answer_category": null
                }
              ]
            }
          ],
          "is_impossible": false
        }
      ],
      "context": "Blok Sensus merupakan Bagian dari suatu wilayah desa/kelurahan yang merupakan daerah kerja dari seorang pencacah. Kriteria blok sensus adalah sebagai berikut:\na. Setiap wilayah desa/kelurahan dibagi habis menjadi beberapa blok sensus.\nb. Blok sensus harus mempunyai batas-batas yang jelas/mudah dikenali, baik batas alam maupun buatan. Batas satuan lingkungan setempat (SLS), seperti: RT, RW, dusun, lingkungan, dan sebagainya diutamakan sebagai batas blok sensus bila batas SLS tersebut jelas (batas alam atau buatan).\nc. Satu blok sensus harus terletak dalam satu hamparan.\n",
      "document_id": 393943
    }
  ],
}

```

**Figure 3.** Data annotated knowledge in SQuAD format

### 3.2. Retriever

At this stage two retriever methods used are Bm25 and Dence retriever. More specifically, we use two Dence retriever models, namely **indobenchmark/indobert-large-p2** and **flax-community/indonesian-roberta-large**. Both models are pretrained language models in Indonesian taken from Huggingface site. two dence retriever models were each compared with Bm25 to see which method had better evaluation results. Apart from that, each dence model was also paired with a Bm25 to find out how well sparse retriever would be paired with dence retriever. Thus there are five retrievers, each of which is tested on top-k documents (k = 1, 5, 10, and 15). five retrievers are:

- Bm25
- indobert-large-p2
- indonesian-roberta-large
- indobert-large-p2 + Bm25
- indonesian-roberta-large + Bm25

### 3.3. Reader

Similar to retriever phase, in this phase a comparison and evaluation will also be carried out on several reader models based on top-k documents (k = 1, 5, 10, and 15). evaluation aims to obtain a reader model based on F1 score and EM metrics. reader model is a pretrained language model in Indonesian taken from Haystack. Reader Model used in this research is:

- esakrissa/IndoBERT-SQuAD
- asaduas/distilbert-base-uncased-indonesia-squadv2
- asaduas/all-MiniLM-L6-v2-indonesia-squadv2
- rizquuula/RoBERTa-IndoSQuADv2\_1691592486-16-2e-05-0.01-5

### 3.4. Experiment scenario

We conducted experiment described in two stages. first stage is to conduct experiments and test performance on five predetermined retriever models. measures used are Recall, Precision, and Mean Average Precision (MAP). From evaluation results, the best retriever will be obtained which will then be used in second stage by pairing it with five previously determined Reader models. measures used are F1-Score and Exact Match



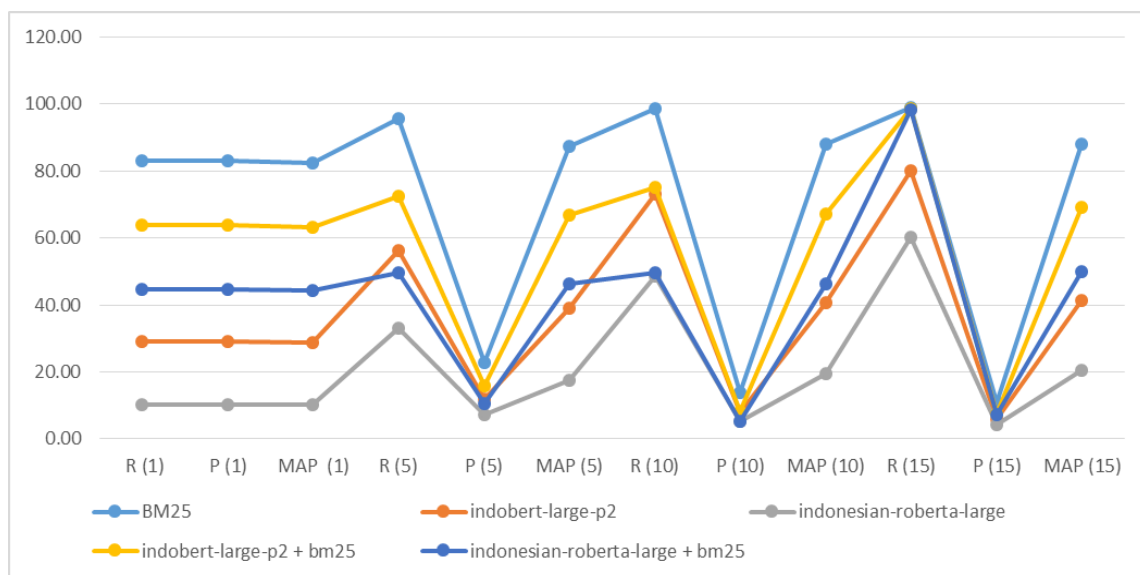
#### 4. Result

The first stage of evaluation can be seen in table 1. Bm25 has highest Recall value among others, namely 98.99% in top-15. Likewise, Precision and MAP have the best scores in top-15. Meanwhile, Indobert-large-p2 and Indonesian-Roberta-large models based on dense retriever actually had worst evaluation results. However, if two models are combined with Bm25, evaluation results are better, although not better than Bm25 alone.

In graph shown in Figure 4, Bm25 has the best performance from top-1 to top 15. This indicates that Bm25 has better performance than Dense Retriever and can even improve performance of Dense Retriever in this study. So in stage two retriever model used is Bm25.

**Table 1.** Evaluation of retriever in first stage

Retriever	top-1			top-5			top-10			top-15		
	R (%)	P (%)	MAP (%)	R (%)	P (%)	MAP (%)	R (%)	P (%)	MAP (%)	R (%)	P (%)	MAP (%)
<b>BM25</b>	82.94	82.94	82.27	95.65	22.74	87.47	98.66	13.83	87.99	98.99	10.93	88.01
<b>indobert-large-p2</b>	29.09	29.09	28.76	56.18	12.04	39.02	73.24	8.06	40.82	80.26	5.90	41.35
<b>indonesian-roberta-large</b>	10.36	10.36	10.20	33.11	7.22	17.65	48.49	5.25	19.51	60.20	4.34	20.39
<b>indobert-large-p2 + bm25</b>	63.87	63.87	63.21	72.57	15.91	66.83	75.25	8.26	67.27	98.66	7.58	69.25
<b>indonesian-roberta-large + bm25</b>	44.81	44.81	44.31	49.49	10.70	46.45	49.49	5.35	46.45	98.32	7.35	49.99



**Figure 4.** Evaluation of retriever in first stage

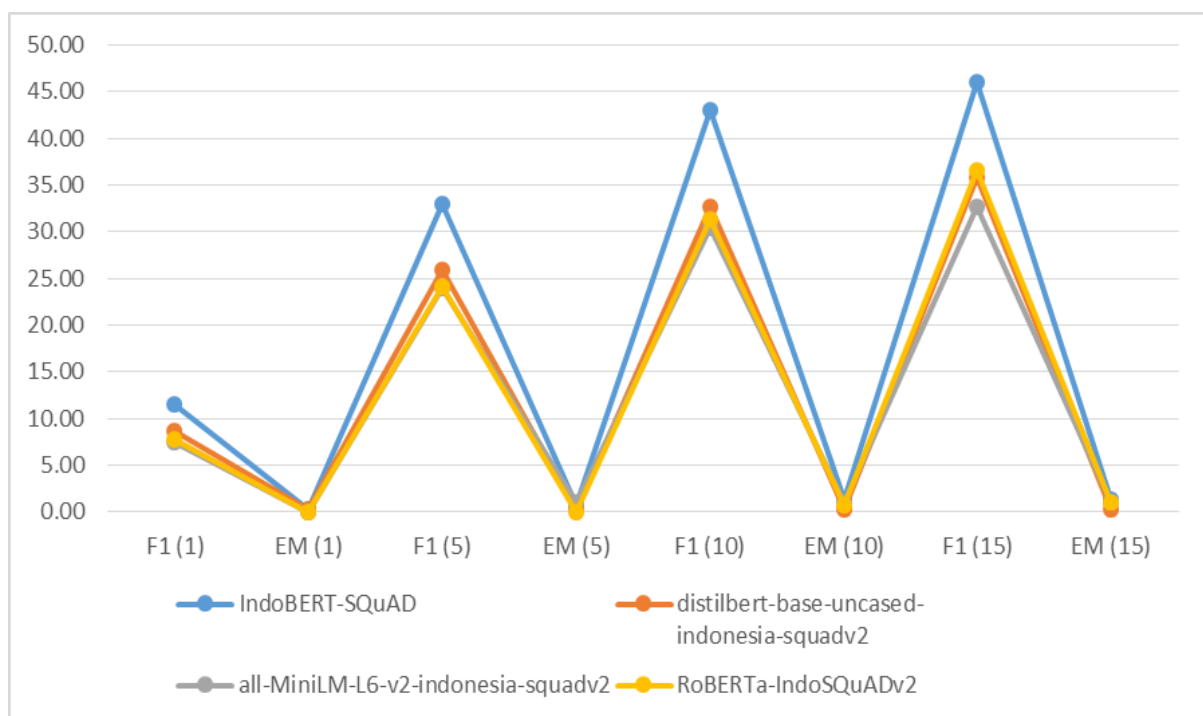
The second stage is an attempt to find out the best combination between retriever and reader. It can be seen from evaluation results that IndoBERT-SQuAD has highest F1 and EM scores compared to



other models in top-15. F1 score is 46.03% while EM score is 1.33%. closest model is RoBERTa-IndoSQuADv2, namely F1 of 36.63% and EM of 1.00%. It can be concluded that IndoBERT-SQuAD has the best performance as shown in Figure 5 which shows IndoBERT-SQuAD has the best F1 score from top-1 to top-15 documents. Even though IndoBERT-SQuAD has the best results, these results are still not good. This result is normal because reader model used did not undergo finetuning on dataset. So that existing reader model does not understand dataset used.

**Table 2.** Evaluate Reader model in second stage

Reader	top-1		top-5		top-10		top-15	
	F1 (%)	EM (%)	F1 (%)	EM (%)	F1 (%)	EM (%)	F1 (%)	EM (%)
<b>IndoBERT-SQuAD</b>	11.55	0.33	33.04	0.66	42.95	1.33	46.03	1.33
<b>distilbert-base-uncased-indonesia-squadv2</b>	8.66	0.33	25.98	0.33	32.66	0.33	35.81	0.33
<b>all-MiniLM-L6-v2-indonesia-squadv2</b>	7.46	0.00	23.94	1.00	30.36	1.00	32.67	1.00
<b>RoBERTa-IndoSQuADv2</b>	7.74	0.00	24.10	0.00	31.30	0.66	36.63	1.00



**Figure 5.** Evaluation Reader model in second stage

## 5. Conclusion

The experimental results show that Bm25 and IndoBERT-SQuAD show the best performance as Retriever-Readers. Where ES is able to retrieve relevant documents by 98.99% in top-15. Meanwhile, retriever model which only consists of Dence retriever actually has worst evaluation results. As for



Reader task, minimum-IndoBERT-SQuAD has the best f1 score of 46.03%. For this reason, this research proposes Bm25 as a Retriever and IndoBERT-SQuAD as a Reader to solve problem in this research. So it can be easier to find appropriate answers to questions on long and complex BPS knowledge.

## 6. Future Work

In future it will be very interesting to add a summarization task. This task can be added before Retriever task or after Reader task. Re-ranking based on personalization is also an interesting undertaking. Where personalization can be done based on metadata from enumerator. Besides that, it can also optimize performance of Reader model by fine-tuning appropriate dataset.

## References

- [1] C. Monz, From document retrieval to question answering. Institute for Logic, Language and Computation, 2003. [Online]. Available: <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.5.5483rep=rep1type=pdf>
- [2] J. Berant and L. Percy, “Semantic parsing via paraphrasing,” vol. 1. Association for Computational Linguistics), 2014, p. 1415–1425. [Online]. Available: <https://doi.org/10.3115/v1/p14-1133>
- [3] K. Xu, S. Reddy, Y. Feng, S. Huang, and D. Zhao, “Question answering on freebase via relation extraction and textual evidence,” vol. 1. Association for Computational Linguistics), 2016, p. 2326–2336. [Online]. Available: <https://doi.org/10.18653/v1/p16-1220>
- [4] F. Zhu, W. Lei, C. Wang, J. Zheng, S. Poria, and T.-S. Chua, “Retrieving and reading: A comprehensive survey on open-domain question answering,” 2021. [Online]. Available: <https://arxiv.org/abs/2101.00774>
- [5] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W. tau Yih, “Dense passage retrieval for open-domain question answering.” Association for Computational Linguistics), 2020, p. 6769–6781. [Online]. Available: <https://doi.org/10.18653/v1/2020.emnlp-main.550>
- [6] Z. H. Syed, A. Trabelsi, E. Helbert, V. Bailleau, and C. Muths, “Question answering chatbot for troubleshooting queries based on transfer learning,” vol. 192. Procedia Computer Science, 2021, p. 941–950. [Online]. Available: <https://doi.org/10.1016/j.procs.2021.08.097>
- [7] R. Anantha, S. Vakulenko, Z. Tu, S. Longpre, S. Stephen, and S. Chappidi, “Open-domain question answering goes conversational via question rewriting.” Association for Computational Linguistics, 2021, p. 520–534. [Online]. Available: <https://doi.org/10.18653/v1/2021.naacl-main.44>
- [8] Q. Zhang, S. Chen, D. Xu, Q. Cao, X. Chen, T. Cohn, and M. Fang, “A survey for efficient open domain question answering,” in Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023, A. Rogers, J. L. Boyd-Graber, and N. Okazaki, Eds. Association for Computational Linguistics, 2023, pp. 14 447–14 465. [Online]. Available: <https://doi.org/10.18653/v1/2023.acl-long.808>
- [9] M. Zhou, Z. Shi, M. Huang, and X. Zhu, “Knowledge- aided open-domain question answering,” 2020. [Online]. Available: <https://arxiv.org/abs/2006.05244>
- [10] F. Zhu, W. Lei, C. Wang, J. Zheng, S. Poria, and T. Chua, “Retrieving and reading: A comprehensive survey on open-domain question answering,” CoRR, vol. abs/2101.00774, 2021. [Online]. Available: <https://arxiv.org/abs/2101.00774>