Common retrieval system evaluation metrics:

Precision:

The ratio of relevant documents over all retrieved documents.

Precision = | { Relevant documents } ∩ { Retrieved documents } | / | { Relevant documents } |

Recall:

Recall is the fraction of the documents that are relevant to the query that are successfully retrieved.

Recall = | { Relevant documents } ∩ { Retrieved documents } | / | { (All) Relevant documents } |

Comprehensive evaluation metrics:

F-score:

The weighted [harmonic mean](https://en.wikipedia.org/wiki/Harmonic_mean" \o "Harmonic mean) of precision and recall, the traditional F-measure or balanced F-score is:

F-score = 2 \* ( precision \* recall ) / ( precision + recall )

MAP(Mean Average Precision): If there are multi retrievals, MAP considers the average precision of all retrievers, usually used in ranking tasks.

Such as some easy tasks as reasoning based on some given contexts, mostly the contexts are very limited and easy to reason. So the following work , the **hard** way, is optimize on how to separate the big data source(here is dpr\_wiki) paragraphs and returned by the retriever(here is dpr(dense passage retriever)) to make the bot get good answers based on the separated contexts.

For the evaluation of a Retriever, the most common one is top k documents hits accuracy. It means when the retriever retrieved top k most similar passages, then apply some metrics e.g. matching of strings, can be recognized as hit. Then let these top k hits divide by the count of altogether questions, that is: top document hit accuracy.

**My time is very limited, I have to work at daytime. I established the environment and make the code runnable. My resource is limited, so I only tried only a very small subset of dpr\_wiki source set, so accuracy seems to be very low.(我是使用我们公司的机器做的，但这上面分配给我的空间只有几十G，还要放代码和checkpoint还有生成的resource文件，所以dpr\_wiki的ctx source set 我只够用很小一部分)**

I use **dpr\_wiki** as my my ctx source and the retriever I used is DPR.

For this problem of split of paragraphs, I have implemented the following 4 algorithms:

1. Fixed length of split of paragraphs. (see **split\_fixed\_length\_of\_words\_with\_sliding\_window.py** in wikiextractor/wikiextractor directory)
2. Slide window split of paragraphs. (see **split\_fixed\_length\_of\_words\_with\_sliding\_window.py** in wikiextractor/wikiextractor directory)
3. Split based on approximate length and sentence boundary detection of paragraphs. (see **passages\_sentence\_boundary\_detection\_parallel.py** in wikiextractor/wikiextractor directory)
4. +Knowledge graph based unrelated passages exclusion. (see **question\_relevancy.py** in wikiextractor directory and **qa\_validation\_knowledge\_graph.py(mind that this is modification to the original file, where the modification is marked the line: “##### Added by Ngaiman Chow for Knowledge Graph”)** in DPR/dpr/data directory)

Code has been in the attachment.

For 1. fixed length of split of paragraphs, I found when the length getting larger and larger, the accuracy also increases. I have tested the fixed lengths as 50 words, 100 words and 200 words.

The top 100 accuracies for fixed length split of paragraphs are 46%, 54% and 62%. And I found the top k accuracy is very low when the k is very small. I’m sorry that I don’t have much time to test since I have to work in the daytime. I believe if I increase the fixed length even more, I will get even better accuracies.

I believe it is simply because of three factors:

①More Information: Longer passages contain more information, which increases the chance that the passage contains the answer to a given query or is relevant to the query.

②Better Context: Longer passages provide more context, which can help the model to better understand the meaning of the text. This is particularly important for models based on transformers like BERT, which rely on context to understand the meaning of words.

③Reduced Fragmentation: If the content is split into smaller passages , information might be fragmented across multiple passages. This can make it harder for the model to find the answer to a query, particularly if the answer depends on information that is spread across multiple passages.

However, there's a trade-off. While longer passages can improve accuracy, they also increase the computational cost of the model, as the model needs to process more words for each passage. Furthermore, if the passages are too long, they might include unrelated information, which can confuse the model. Therefore, it's important to find a balance between passage length and accuracy.

**Examples:**

1. **For the question 9: when is the last time the philadelphia won the superbowl ?**

**The answer is ['Super Bowl LII ,', '2017']**

**For the fixed length 100 words split , the following line:**

**Giants, the 2016 home opener against the Cleveland Browns, and the 2017 home opener against the Giants. In the 2010 season against the Green Bay Packers, on September 12, 2010, the Eagles wore uniforms similar to the ones that were worn by the 1960 championship team in honor the 50th anniversary of**

**is located in an introduction to the team philadelphia, and longer and ranked higher than(in fixed length 100 words split it is at Top 1, in fixed length 50 words split, only in Top 3 appears a related passage) the fixed length 50 word split.**

1. **For the question 4: the south west wind blows across nigeria between ?**

**The answer is ['till September']**

**For the fixed length 200 words split, the following line:**

**an annual rainfall of about with a single rainfall maxima in September. The single Dry season experienced in this climate, the tropical savanna climate in central Nigeria beginning from December to march, is hot and dry with the Harmattan wind, a continental tropical (CT) airmass laden with dust from the Sahara Desert prevailing throughout this period. With the Intertropical Convergence Zone (ITCZ) swinging northward over West Africa from the Southern Hemisphere in April, heavy showers coming from pre-monsoonal convective clouds mainly in the form of squall lines also known as the north easterlies formed mainly as a result of the interactions of the two dominant airmasses in Nigeria known as the Maritime tropical(south westerlies) and the Continental tropical(north easterlies), begins in central Nigeria while the Monsoons from the south atlantic ocean arrives in central Nigeria in July bringing with it high humidity, heavy cloud cover and heavy rainfall which can be daily occurrence lasting till September when the monsoons gradually begin retreating southward to the southern part of Nigeria.Rainfall totals in central Nigeria varies from in the lowlands of the river Niger Benue'",**

**is located in an introduction to Nigeria, and longer and ranked higher than(in fixed length 200 words split it is at Top 1, in fixed length 100 words split, only Top 4 appears a related passage) the fixed length 100 words split.**

For 2. slide window split of paragraphs, I set the fixed length still as 50 words, 100 words and 200 words. And the sliding window as 10 words, 20 words and 40 words respectively. I have not tested other configurations basically because I don’t have much time I’m sorry, I have to work in daytime.

In theory, applying a sliding window split rather than the original fixed length split can have several effects with its benefits:

1. Overlap of Context: Sliding window split creates overlapping passages, which means that some information will be present in multiple passages. This can be beneficial in some cases, as it ensures that important context is not lost at the boundaries of the passages.

2. Increased Number of Passages: Sliding window split increases the total number of passages, as it creates a new passage for every step of the window. This can increase the coverage of the text and the chances of finding relevant passages, but it also increases the computational cost of the retrieval process.

3. Potential Improvement in Retrieval Performance: By preserving more context and increasing the coverage of the text, sliding window split can potentially improve the performance of the retrieval process. However, this is not guaranteed and depends on the specific task and dataset.

4. Potential Increase in Redundancy: As sliding window split creates overlapping passages, it can increase the redundancy in the retrieved passages. This means that the top-k retrieved passages might contain duplicate or very similar information.

In summary, sliding window split can provide more context and coverage than fixed length split, but it can also increase redundancy and computational cost. The impact on retrieval performance can vary depending on the specific task and dataset.

But for me, in this my practice, it factually decreases accuracy, but only a little. The top 100 accuracies for fixed length 50 split with sliding window 10, fixed length 100 with sliding window 20 and fixed length with sliding window 40 of paragraphs are 45%, 53% and 61% respectively. I believe it is mainly because of two reasons. The first reason is that when using sliding window the context source set becomes large. The second reason is that in my case, since when the k is very small the top k accuracy is very low means that most retrieved results in the top rank are not related. By doing so, these unrelated results and their similar sibling results(by sliding window) tend to co-exist in the results by sliding windows split which took the positions that should be for the related retrieval results.

**Examples:**

**Although for some question fixed length 100 split with sliding window 20 produces better results than fixed length 100 split only, but the majority is poorer:**

1. **For the question 12: swan lake the sleeping beauty and the nutcracker are three famous ballets by?**

**Answer: ['Pyotr Ilyich Tchaikovsky']**

**For fixed length 100 words split, the following passage which has something related to introduction of Pyotr Ilyich Tchaikovsky:**

**E. T. A. Hoffmann Ernst Theodor Amadeus Hoffmann (commonly abbreviated as E. T. A. Hoffmann; born Ernst Theodor Wilhelm Hoffmann; 24 January 177625 June 1822) was a German Romantic author of fantasy and Gothic horror, a jurist, composer, music critic, draftsman and caricaturist. His stories form the basis of Jacques Offenbach\\'s famous opera \"The Tales of Hoffmann\", in which Hoffmann appears (heavily fictionalized) as the hero. He is also the author of the novella \"The Nutcracker and the Mouse King\", on which Pyotr Ilyich Tchaikovsky\\'s ballet \"The Nutcracker\" is based. The ballet \"Copp\u00e9lia\" is'",**

**with some other passages which are more than and ranked higher(about Top 30) than fixed length 100 with sliding window 20 split.**

1. **For the question 15: who designed the garden city of new earswick?**

**Answer: ['Raymond Unwin', 'architect Barry Parker', 'planner Raymond Unwin']**

**For fixed length 100 words split, the following passage which has introduction something related to Raymond Unwin:**

**Green and created the Hampstead Garden Suburb. The suburb attracted the talents of architects including Raymond Unwin and Sir Edwin Lutyens, and it ultimately grew to encompass over 800 acres. During the First World War the Tudor Walters Committee was commissioned to make recommendations for the post war reconstruction and housebuilding. In part, this was a response to the shocking lack of fitness amongst many recruits during World War One, attributed to poor living conditions; a belief summed up in a housing poster of the period \"you cannot expect to get an A1 population out of C3 homes\“,**

**which ranked higher than fixed length 100 words split with sliding window 20 words(for the former it ranked at top 3, for the later it ranked at top 6 which supports my guess that took the positions that should be for the related retrieval results)**

For 3. Split based on approximate length and sentence boundary detection of paragraphs, I use multiprocessing to boost the speed of sentence boundary detection. In thoery, by using sentence boundary detection split can we receive the following benefits:

1.Preserving Semantic Units: Sentences are natural semantic units in language. By splitting text at sentence boundaries, we ensure that these semantic units are preserved, which can help to maintain the meaning and context of the text.  
2. Improving Model Performance: Many NLP models, such as transformers and recurrent neural networks, work better with shorter sequences of text. By splitting text into sentences, we can feed these models with shorter sequences, which can help to improve their performance.  
3. Reducing Fragmentation: Sentence boundary detection can help to reduce the fragmentation of information. If text is split at arbitrary points, a sentence might be split across two passages, which can disrupt the flow of information.  
However, there can also be some drawbacks:  
1. Variable Length: Sentences can vary greatly in length. This can result in passages of very different lengths, which can be challenging to handle in some NLP models that require fixed-length input.  
2. Loss of Context: If sentences are treated as separate passages, the context between sentences can be lost. This can be problematic for models that rely on context to understand the meaning of the text.  
3. Computational Cost: Sentence boundary detection can be more computationally expensive than fixed-length splitting, as it requires additional processing to identify the sentence boundaries.  
  
In summary, sentence boundary detection can help to preserve semantic units and improve model performance, but it can also result in variable length passages and potentially increase computational cost.

By using sentence boundary detection, I did get much more satisfying results, for every corresponding configurations that does not applying sentence boundary detection, the sentence boundary detection counterpart almost increased top 100 accuracy by more 5% to, i.e. 50%, 59% and 66%.

**Examples:**

1. **For the question 24: what part of the brain is in the middle cranial fossa ?**

**Answer: ['pituitary gland', 'the temporal lobes', 'temporal lobes', 'the pituitary gland']**

**for the sentence-boundary-detection model the passage:**

**Pituitary gland In vertebrate anatomy, the pituitary gland, or hypophysis, is an endocrine gland about the size of a pea and weighing in humans. It is a protrusion off the bottom of the hypothalamus at the base of the brain. The hypophysis rests upon the hypophysial fossa of the sphenoid bone in the center of the middle cranial fossa and is surrounded by a small bony cavity (sella turcica) covered by a dural fold (diaphragma sellae). The anterior pituitary (or adenohypophysis) is a lobe of the gland that regulates several physiological processes (including stress, growth, reproduction, and lactation).**

**where the answer is located almost at the start of a sentence, has been retrieved at the first being a top 3 results, while for the non-sentence-boundary-detection model on retrieved the top ranked passage:**

**function of sorting information from the periphery, the thalamus also connects the cerebellum and basal ganglia with the cerebrum. In common with the aforementioned reticular system the thalamus is involved in wakefullness and consciousness, such as though the SCN. The hypothalamus engages in functions of a number of primitive emotions or feelings such as hunger, thirst and maternal bonding. This is regulated partly through control of secretion of hormones from the pituitary gland and additionally the hypothalamus plays a role in motivation and many other behaviors of the individual. The**

**as only at top 24.**

1. **For the question 46: what network is showing the monday night football game?**

**Answer: ['ESPN']**

**for the sentence-boundary-detection model the passage:**

**The league also provides RedZone, an omnibus telecast that cuts to the most relevant plays in each game, live as they happen. In addition to the regional games, the league also has packages of telecasts, mostly in prime time, that are carried nationwide. NBC broadcasts the primetime \"Sunday Night Football\" package', which includes the Thursday NFL Kickoff game that starts the regular season and a primetime Thanksgiving Day game. ESPN carries all Monday Night Football games. The NFL's own network, NFL Network, broadcasts a series titled \"Thursday Night Football\", which was originally exclusive to the network, but which in recent years has had several games simulcast on CBS (since 2014) and NBC (since 2016) (except the Thanksgiving and kickoff games, which remain exclusive to NBC).",**

**where the answer is located at the start of some sentence, has been retrieved at the first being a top 1 result, and with others, they are the results somewhat more than those retrieved by the non-sentence-boundary-detection model.**

For 4. Knowledge graph based unrelated passages exclusion, since my time is limited, I only implemented a method of question relevancy detection by an online requesting to “https://api.conceptnet.io/relatedness?node1=/c/en/{}&node2=/c/en/{}” . I also use gensim to get the keywords for passages and questions.

**Examples:**

1. **question 15:who designed the garden city of new earswick?**

**The detected Keywords: [ 'garden', 'earswick' ]**

**The retrieved passage for question is not relevant(False):**

**Horace Walpole Horatio Walpole, 4th Earl of Orford (; 24 September 1717 \u2013 2 March 1797), also known as Horace Walpole, was an English art historian, man of letters, antiquarian and Whig politician. He had Strawberry Hill House built in Twickenham, south-west London, reviving the Gothic style some decades before his Victorian successors. His literary reputation rests on the first Gothic novel, \"The Castle of Otranto\" (1764), and his \"Letters\", which are of significant social and political interest. They have been published by Yale University Press in 48 volumes. He was the son of the first'**

**The detected keywords for this passage is Keywords: ['gothic', 'september', 'walpole', 'university', 'reputation', 'hill', 'house', 'art', 'historian', 'london', 'south', 'politician', 'whig', 'victorian', 'antiquarian', 'successors', 'significant']**

**The knowledge graph method of question relevance: False**

**So this unrelevant passage is excluded by the knowledge graph method. A good sample!**

**The retrieved passage for question is not relevant(False):**

**Le Corbusier Charles-\u00c9douard Jeanneret (6 October 1887 \u2013 27 August 1965), known as Le Corbusier (), was a Swiss-French architect, designer, painter, urban planner, writer, and one of the pioneers of what is now called modern architecture. He was born in Switzerland and became a French citizen in 1930. His career spanned five decades, and he designed buildings in Europe, Japan, India, and North and South America. Dedicated to providing better living conditions for the residents of crowded cities, Le Corbusier was influential in urban planning, and was a founding member of the Congr\u00e8s'",**

**The detected keywords for this passage is Keywords: ['french', 'designer', 'designed', 'japan', 'america', 'modern', 'painter urban', 'better living', 'august', 'october', 'planner', 'architect', 'founding member', 'career spanned', 'crowded']**

**The knowledge graph method of question relevance: False**

**So this unrelevant passage is excluded by the knowledge graph method. A good sample!**

**The retrieved passage for question is relevant(True):**

**Green and created the Hampstead Garden Suburb. The suburb attracted the talents of architects including Raymond Unwin and Sir Edwin Lutyens, and it ultimately grew to encompass over 800 acres. During the First World War the Tudor Walters Committee was commissioned to make recommendations for the post war reconstruction and housebuilding. In part, this was a response to the shocking lack of fitness amongst many recruits during World War One, attributed to poor living conditions; a belief summed up in a housing poster of the period \"you cannot expect to get an A1 population out of C3 homes\" -'",**

**The detected keywords for this passage is Keywords: ['war', 'living', 'walters', 'edwin', 'garden', 'suburb', 'raymond', 'including', 'ultimately grew', 'housing poster', 'belief summed']**

**The knowledge graph method of question relevance: True**

**So this relevant passage is not excluded i.e. included by the knowledge graph method. A good sample!**

1. **question 29: in which regions are most of africa petroleum and natural gas found?**

**The detected Keywords: [ 'petroleum', 'nautural', 'gas' ]**

**The retrieved passage for question is not relevant(False):**

**side of the continent nearest its source. The Rovuma, Rufiji and Tana principally drain the outer slopes of the African Great Lakes highlands. In the Horn region to the north, the Jubba and the Shebelle rivers begin in the Ethiopian Highlands. These rivers mainly flow southwards, with the Jubba emptying in the Indian Ocean. The Shebelle River reaches a point to the southwest. After that, it consists of swamps and dry reaches before finally disappearing in the desert terrain near the Jubba River. Another large stream, the Hawash, rising in the Ethiopian mountains, is lost in a saline,**

**The detected keywords for this passage is Keywords: [ ‘africa’, ‘production’, ‘american’, ‘place’, ‘consultancy’, ‘metals’, ‘region’, ‘tanzania’, ‘significant’, ‘south’, ‘namibia’, ‘zimbabwe’, ‘supplier’, ‘major’, ‘supliers’ ]**

**The knowledge graph method of question relevance: False**

**So this unrelevant passage is excluded by the knowledge graph method. A good sample!**

**The retrieved passage for question is relevant(True):**

**containing different vegetation from the salt water mangrove swamps, and north of that is rain forest. The savannah zone's three categories are divided into mm' Guinean forest-savanna mosaic, made up of plains of tall grass which are interrupted by trees, the most common across the country; Sudan savannah, similar but with shorter grasses and shorter trees; and Sahel savannah patches of grass and sand, found in the northeast. Nigeria's natural resources include but are not limited to petroleum (see petroleum in Nigeria), tin, columbite, iron ore, coal, limestone, lead, zinc, natural gas,**

**The detected keywords for this passage is Keywords: [ ‘savannah’, ‘forest’, ‘different’, ‘grass’, ‘sudan’, ‘tin’, ‘zinc’, ‘lead’, ‘ore’, ‘water’, ‘mangrove’, ‘shorter’, ‘grasses’, ‘nigera’, ‘natural’, ‘resources’, ‘columbite’, ‘iron’ ]**

**The knowledge graph method of question relevance: True**

**So this relevant passage is not excluded i.e. included by the knowledge graph method. A good sample!**

Can test the whole effects of separation effects with the retriever by using the two files: **rag\_test.py** and **rag\_test\_knowledge\_graph.py** in DPR directory. The later is an variant of the former but added with optimization with knowledge graph, their parameters are can be found in launch.json, the later could also tune two extra params: top\_k and relevancy\_threshold.