Extrinsic evaluation of question generation methods with user journey logs

Elie Antoine[1], Eléonore Besnehard[2],Frédéric Béchet[1,5], Géraldine Damnati[3],
Eric Kergosien[2], Arnaud Laborderie[4]
(1) LIS, Aix Marseille Univ; (2) Geriico, Université de Lille; (3) Orange Innovation;
(4) BNF - Bibliothèque Nationale de France
(5) International Laboratory on Learning Systems (ILLS - IRL CNRS), Montreal

Abstract

There is often a significant disparity between the performance of Natural Language Processing (NLP) tools as evaluated on benchmark datasets using metrics like ROUGE or BLEU, and the actual user experience encountered when employing these tools in real-world scenarios. This highlights the critical necessity for user-oriented studies aimed at evaluating user experience concerning the effectiveness of developed methodologies. A primary challenge in such ”ecological” user studies is their assessment of specific configurations of NLP tools, making replication under identical conditions impractical. Consequently, their utility is limited for the automated evaluation and comparison of different configurations of the same tool. The objective of this study is to conduct an extrinsic evaluation of a question generation system within the context of an external task involving document linking. To do this we conducted an ”*ecological*” evaluation of a document linking tool in the context of the exploration of a Social Science archives and from this evaluation, we aim to derive a form of a ”*reference corpus*” that can be used offline for the automated comparison of models and quantitative tool assessment. This corpus is available on the following link: `https://gitlab.lis-lab.fr/archival-public/autogestion-qa-linking`

1. Introduction

Question Generation (QG) from text is a key task in Natural Language Processing (NLP), attracting increased attention for its role in testing the syntactic and semantic understanding of generative language models. Recent literature, including Guo et al. (2024), documents the development and comparison of various neural generation techniques, benchmarked against datasets like SQuAD (Rajpurkar et al., 2016) using automated evaluation metrics.

Intrinsic evaluations compare machine-generated questions with a human-produced reference set, employing ngram-based metrics such as ROUGE to measure text fluency and semantic similarity metrics like BERTScore, which uses pre-trained BERT embeddings and cosine similarity to assess the closeness of machine and human-generated text.

Additionally, the relevance of human evaluations in assessing question quality is crucial. For example, Bojic et al. (2023) proposes a hierarchical set of criteria for evaluating the semantic content and formulation of Machine Reading Comprehension input questions. As discussed by the authors, these *intrinsic* benchmark evaluations primarily assess question quality, they seldom address the ”*usefulness*” of questions in specific applications, indicating a need for *extrinsic* evaluation methods.

We propose to evaluate question generation models through the task of document linking in the general context of exploring Social Science archives with specialized users.

Document Linking consists in adding hyperlinks between documents of a collection according to some criteria. When the criteria are explicit, like in Wikipedia, evaluating the relevance of predicted links means comparing them to a reference containing explicit hyperlinks (Brochier and Béchet, 2021). However when the links are implicit, which is the case when dealing with linking criteria such as *textual similarity* or *entailment*, evaluating links relevance becomes difficult as it relies on subjective criteria and therefore collecting *gold annotation* on such data is a challenge.

In this study, we chose to conduct an experiment using user journey logs [1] to establish sets of related documents during a session. These sets can be used to compare and evaluate different question generation system through the link they produce. Specifically, our aim is to evaluate our question generation system with the ”*question-linking*” paradigm as presented in Antoine et al. (2023) with real users by observing the journeys of a panel of testers. These testers explored an interface designed for discovering a collection of journal archives, which offered various exploration options. Among these options, users could select a passage and open a window containing linked passages from other articles.

We compares the links produced by four strategies: the first uses paragraph similarity as a baseline, the second involves similarity between (ques-

[1] The corpus collected in this study is available on the following link : `https://gitlab.lis-lab.fr/archival-public/autogestion-qa-linking`

Proceedings of the HumEval @LREC-COLING 2024, pages 63–70
Turin, Italy. May 21st, 2024. ©2024 European Language Resources Association (ELRA)

tion, answer) pairs, with questions generated by a small (< 1B) model and answers extracted from the text. The last two strategies focus solely on question similarity, with one employing a small (< 1B) language model and the other a large (7B) Language Model (LLM).

2. Question generation for exploring archive collections

We have explored the potential utility of Question Generation models in the context of exploring a collection of documents. Even if current Question-Answer (QA) models might be too simplistic for use in practical archive exploration, the focus here is on the use of Question Generation models. These models are trained differently from QA models, as they are designed to predict a question based on an answer and a text document, as opposed to generating a response given a question and a document.

The key idea is to use Question Generation models to characterize documents in archives by creating a set of questions associated with the text segments. This is achieved by selecting potential answers from text segments and generating questions based on these answers and their context. By comparing the questions and answers from different documents, the system can predict links between them, effectively adding an *explainability layer* to the document exploration process. This allows users to quickly assess the relevance of links by examining the associated QA pairs, which can save time compared to the traditional approach of following every link to determine its significance.

We present below a short description of our question generation and linking methods.

Question generation methods

We automatically generated questions on the collection using the same method as the one described in Antoine et al. (2023). In this approach, a semantic parser is used to select potential answers from the articles. As proposed in Pyatkin et al. (2021) and Bechet et al. (2022), a Semantic Role Labelling (SRL) model following the PropBank formalism (Palmer et al., 2005) is used in order to select answers candidates among the detected semantic roles. Following this step, a question generation model is used to provide a question, given the selected answer and its context. This model is trained by fine-tuning the BARThez (Kamal Eddine et al., 2021) language model on a French corpus of question-answer-context triplets called *FQuAD* (d'Hoffschmidt et al., 2020). To address the model's tendency to overgenerate potentially meaningless or overly simplistic questions, a series of filters are then applied to enhance quality and reduce quantity.These filters are based on resources such as a thesaurus or a list of persons linked to the applicative domains as well as textual indicators. Here is a list of the indicators considered in the filtering process:

1. #$(pers)$: the number of person mentions belonging to a given list
2. #(th_answer) and #$(th\_question)$: the number of keywords from the thesaurus of notions in respectively the answer and the question (to add a control on the semantic relevance of the question)
3. We compute the average length of the generated questions and extracted answers to calculate the deviation from the mean of each questions $(quest\_diff\_mean)$ and answers (ans_diff_mean)
4. We finally compute $inter\_qa$, the percent of intersection between the extracted answer and the question (to avoid nonsensical questions that contain the answer to their own question).

All these filters are used in a decision rule that accept or reject a generated pair question/answer.

Linking methods

Links between items in the collection are produced using the same method as in Antoine et al. (2023). The proposed approach is to generate links using questions and answers generated from the text rather than directly on the text itself. The embedding projection for each "`<question> | <answer>`" pair structure uses the Sentence-Transformer (Reimers and Gurevych, 2019) library, and more precisely the multilingual model *distiluse-base-multilingual-cased-v1* (Reimers and Gurevych, 2020). A cosine similarity measure is then employed between all pairwise combinations of these embeddings, resulting in the computation of a similarity matrix.

In this study we will perform an extrinsic human evaluation where the usefulness of the questions for document linking is studied.

3. Collecting logs from an exploration interface

This study was conducted within the framework of the French ANR project ARCHIVAL[2], aimed at developing novel exploration methods for thematic archive collections using machine comprehension techniques. The archive collection chosen for this study is a collection of social science journal articles

[2] https://anr.fr/Projet-ANR-19-CE38-0011

in French from the *Autogestion* (Self-management) journal[3]. This collection is distributed in its digitized form by the French Persée organization. It is part of a larger pluridisciplinary multilingual mixed collection (archives and documents) that has been gathered since the 1960's by the FMSH[4] foundation's library. The full collection has been granted the Collex label (*Collection d'Excellence* or Excellency Collection) from the CollEx-Persée[5] network under the supervision of higher education and research for the preservation of corpus of digitized or natively digital documents.

This collection, published during a period ranging from the 1960s to the 1980s, constitute a corpus of 46 issues for an overall amount of 896 articles (more than 6000 pages and 1.98M tokens).

In order to navigate in this large collection, the interface homepage proposes a search engine and two main access modes: a direct access through timelines, tables of content and indexes containing references to persons (all the authors and people mentioned in the documents content) and notions from a thesaurus of around 400 notions specifically designed for the semantic domain of the journal. The notions and persons are automatically detected from the text of the articles, the method and the exploitation of these functionalities by users are out of the scope of the current study and are not detailed here.

Once a user has entered the collection and opened an article, he can further explore it with linking mechanisms. The user can select a text area in the article which becomes highlighted. This selection corresponds to a particular area of interest for which links to other documents in the collection can be proposed according to two methodologies:

- Firstly question-linking method presented in the previous section. A list of questions generated from the paragraph containing the highlighted text is displayed to the user who can click on any of these questions to obtain a list of n links to related paragraphs in other documents calculated thanks to the method presented before. Links are associated here to references to the title and the authors of the target documents as well as a snippet of the target paragraph. The link is explained by the pair of questions from the source and the target paragraph. An example of document linking and question explanation is given in figure 1.
- Secondly, a method based on textual similarity using *SentenceBert* (Reimers and Gurevych, 2019) is applied to the paragraph containing the highlighted text in order to display the n other paragraphs in the collection that minimize the similarity criteria.

In our experiments, the amount of displayed links was set to $n = 10$. An illustration of this text selection and linking presentation method is given in figure 2. Users can choose the document linking method they want to use.

Within the interface, users can perform a variety of actions to navigate and manipulate content. First, they can open or close windows associated with articles, notions, or persons.They can also switch between different views, including the timeline, notions page, and persons page.They can switch submenus within an article window, whether it's toggling between viewing the article text, the notions automatically extracted from the text or the person cited in the text.
To seek relevant connections, users can also use the links provided when selecting an area of interest. All these actions are logged.

A first way of exploiting the logs would be to analyse if users actually clicked on the links proposed by the various algorithms. If this is an interesting way to analyze user journeys and their acceptation of the functionalities, it is not enough to provide a reproducible evaluation framework to compare several question generation approaches or several linking strategies. In this work, in order to propose a reproducible evaluation protocole, we consider that the set of documents consulted by a given user during a test session constitutes a coherent set of documents that are of interest for him/her. We will call this set of consulted documents a *user-log collection*. Then we want to check a posteriori if, starting from one document of the collection, a given exploration approach would allow to reach other documents from the same collection. We formulate the hypothesis that proposing links that allow users to reach more easily other documents of interest is more helpful. Hence we can compare several linking methods, beyond the ones that were originally implemented during the collection phase.

4. From log collection to extrinsic evaluation

This section describes how we turn the set of documents in our corpus into a graph according to a given linking method, and how we can evaluate such graphs thanks to the user-log collections described earlier.

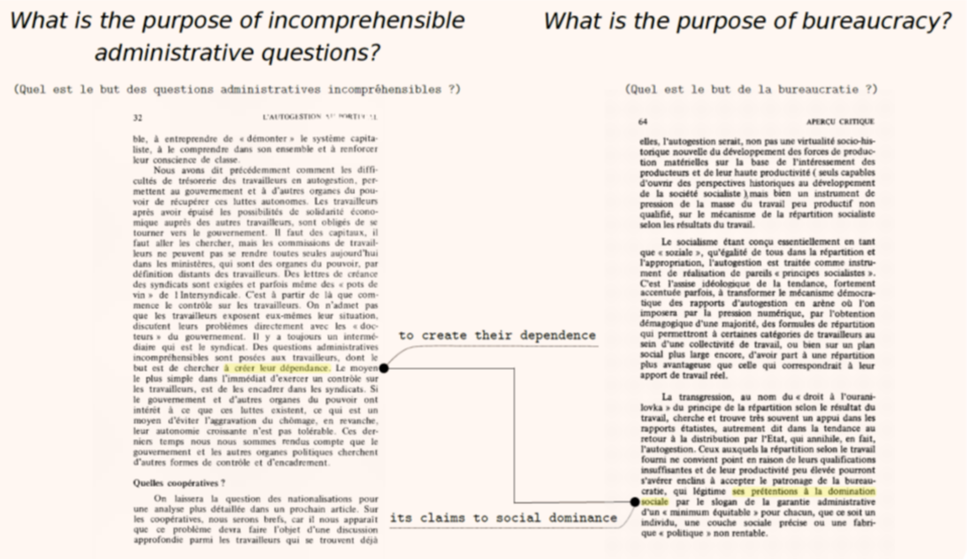
Graph creation

For each linking method L, the first step in our process is to turn our document collection into a

[3] https://www.persee.fr/collection/autog

[4] Fondation Maison des Sciences de l'Homme, https://www.fmsh.fr/

[5] https://www.collexpersee.eu/le-reseau/



What is the purpose of incomprehensible administrative questions?

(Quel est le but des questions administratives incompréhensibles ?)

32 L'AUTOGESTION AU PORTUGAL

ble, à entreprendre de « démonter » le système capitaliste, à le comprendre dans son ensemble et à renforcer leur conscience de classe.

Nous avons dit précédemment comment les difficultés de trésorerie des travailleurs en autogestion, permettent au gouvernement et à d'autres organes du pouvoir de récupérer ces luttes autonomes. Les travailleurs après avoir épuisé les possibilités de solidarité économique auprès des autres travailleurs, sont obligés de se tourner vers le gouvernement. Il faut des capitaux, il faut aller les chercher, mais les commissions de travailleurs ne peuvent pas se rendre toutes seules aujourd'hui dans les ministères, qui sont des organes du pouvoir, par définition distants des travailleurs. Des lettres de créance des syndicats sont exigées et parfois même des « pots de vin » de 1 Intersyndicale. C'est à partir de là que commence le contrôle sur les travailleurs. On n'admet pas que les travailleurs exposent eux-mêmes leur situation, discutent leurs problèmes directement avec les « docteurs » du gouvernement. Il y a toujours un intermédiaire qui est le syndicat. Des questions administratives incompréhensibles sont posées aux travailleurs, dont le but est de chercher à créer leur dépendance. Le moyen le plus simple dans l'immédiat d'exercer un contrôle sur les travailleurs, est de les encadrer dans les syndicats. Si le gouvernement et d'autres organes du pouvoir ont intérêt à ce que ces luttes existent, ce qui est un moyen d'éviter l'aggravation du chômage, en revanche, leur autonomie croissante n'est pas tolérable. Ces derniers temps nous nous sommes rendus compte que le gouvernement et les autres organes politiques cherchent d'autres formes de contrôle et d'encadrement.

Quelles coopératives ?

On laissera la question des nationalisations pour une analyse plus détaillée dans un prochain article. Sur les coopératives, nous serons brefs, car il nous apparait que ce problème devra faire l'objet d'une discussion approfondie parmi les travailleurs qui se trouvent déjà

What is the purpose of bureaucracy?

(Quel est le but de la bureaucratie ?)

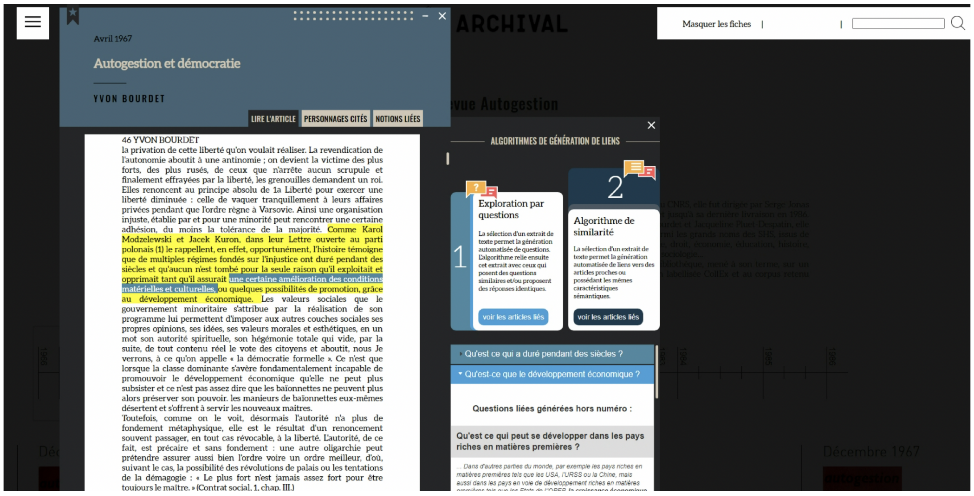
64 APERÇU CRITIQUE

elles, l'autogestion serait, non pas une virtualité socio-historique nouvelle du développement des forces de production matérielles sur la base de l'intéressement des producteurs et de leur haute productivité (seuls capables d'ouvrir des perspectives historiques au développement de la société socialiste), mais bien un instrument de pression de la masse du travail peu productif non qualifié, sur le mécanisme de la répartition socialiste selon les résultats du travail.

Le socialisme étant conçu essentiellement en tant que « soziale », qu'égalité de tous dans la répartition et l'appropriation, l'autogestion est traitée comme instrument de réalisation de pareils « principes socialistes ». C'est l'assise idéologique de la tendance, fortement accentuée parfois, à transformer le mécanisme démocratique des rapports d'autogestion en arène où l'on imposera par la pression numérique, par l'obtention démagogique d'une majorité, des formules de répartition qui permettront à certaines catégories de travailleurs au sein d'une collectivité de travail, ou bien sur un plan social plus large encore, d'avoir part à une répartition plus avantageuse que celle qui correspondrait à leur apport de travail réel.

La transgression, au nom du « droit à l'ouranilovka » du principe de la répartition selon le résultat du travail, cherche et trouve très souvent un appui dans les rapports étatistes, autrement dit dans la tendance au retour à la distribution par l'Etat, qui annihile, en fait, l'autogestion. Ceux auxquels la répartition selon le travail fourni ne convient point en raison de leurs qualifications insuffisantes et de leur productivité peu élevée pourront s'avérer enclins à accepter le patronage de la bureaucratie, qui légitime ses prétentions à la domination sociale par le slogan de la garantie administrative d'un « minimum équitable » pour chacun, que ce soit un individu, une couche sociale précise ou une fabrique « politique » non rentable.

Figure 1: Example of highligthed source and target paragraph with question explanation



Avril 1967

Autogestion et démocratie

YVON BOURDET

46 YVON BOURDET

la privation de cette liberté qu'on voulait réaliser. La revendication de l'autonomie aboutit à une antinomie ; on devient la victime des plus forts, des plus rusés, de ceux que n'arrête aucun scrupule et finalement effrayées par la liberté, les grenouilles demandent un roi. Elles renoncent au principe absolu de 1a Liberté pour exercer une liberté diminuée : celle de vaquer tranquillement à leurs affaires privées pendant que l'ordre règne à Varsovie. Ainsi une organisation injuste, établie par et pour une minorité peut rencontrer une certaine adhésion, du moins la tolérance de la majorité. Comme Karol Modzelewski et Jacek Kuron, dans leur Lettre ouverte au parti polonais (1) le rappellent, en effet, opportunément, l'histoire témoigne que de multiples régimes fondés sur l'injustice ont duré pendant des siècles et qu'aucun n'est tombé pour la seule raison qu'il exploitait et opprimait tant qu'il assurait une certaine amélioration des conditions matérielles et culturelles, ou quelques possibilités de promotion, grâce au développement économique. Les valeurs sociales que le gouvernement minoritaire s'attribue par la réalisation de son programme lui permettent d'imposer aux autres couches sociales ses propres opinions, ses idées, ses valeurs morales et esthétiques, en un mot son autorité spirituelle, son hégémonie totale qui vide, par la suite, de tout contenu réel le vote des citoyens et aboutit, nous Je verrons, à ce qu'on appelle « la démocratie formelle ». Ce n'est que lorsque la classe dominante s'avère fondamentalement incapable de promouvoir le développement économique qu'elle ne peut plus subsister et ce n'est pas assez dire que les baïonnettes ne peuvent plus alors préserver son pouvoir. les manieurs de baïonnettes eux-mêmes désertent et s'offrent à servir les nouveaux maitres.

Toutefois, comme on le voit, désormais l'autorité n'a plus de fondement métaphysique, elle est le résultat d'un renoncement souvent passager, en tout cas révocable, à la liberté. L'autorité, de ce fait, est précaire et sans fondement : une autre oligarchie peut prétendre assurer aussi bien l'ordre voire un ordre meilleur, d'où, suivant le cas, la possibilité des révolutions de palais ou les tentations de la démagogie : « Le plus fort n'est jamais assez fort pour être toujours le maître. » (Contrat social, 1, chap. III.)

Exploration par questions

La sélection d'un extrait de texte permet la génération automatisée de questions. L'algorithme relie ensuite cet extrait avec ceux qui posent des questions similaires et/ou proposent des réponses identiques.

Algorithme de similarité

La sélection d'un extrait de texte permet la génération automatisée de liens vers des articles proches ou possédant les mêmes caractéristiques sémantiques.

- Qu'est ce qui a duré pendant des siècles ?
- Qu'est-ce que le développement économique ?

Questions liées générées hors numéro :

Qu'est ce qui peut se développer dans les pays riches en matières premières ?

... Dans d'autres parties du monde, par exemple les pays riches en matières premières tels que les USA, l'URSS ou la Chine, mais aussi dans les pays en voie de développement riches en matières premières tels que les Etats de l'"OPEP **la croissance économique**

Figure 2: Text selection and link production interface

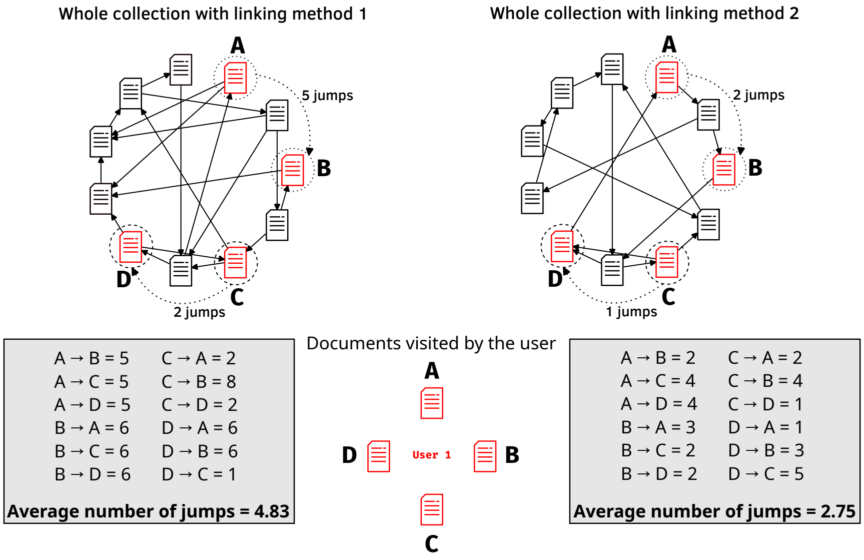
graph G_L. This is achieved by considering all the documents within our corpus which contain automatically generated links. Each node in G_L corresponds to a document (an article of the *Autogestion* journal), and we add an edge between document A and document B, noted (A, B), if there is at least one link connecting a paragraph from A to a paragraph in B thanks to the linking method L. This is a directed graph as all linking methods are not necessarily symmetrical.

We apply a weight to all the (A, B) edges of this graph between a document A and a document B according to the following principle:

1. for each edge (A, B) we compute the number of direct links between documents A and B, called $N_L(A, B)$
2. to normalize these numbers at the document level, for each document A, we rank all the outgoing edges from A to any other document $(A, .)$ in the collection according to the values $N_L(A, .)$.
3. the weight of edge (A, B) called $W_L(A, B)$ is the rank of this edge among all the outgoing edges from document A sorted by $N_L(A, .)$.

The best weight an edge (A, B) can have is $W_L(A, B) = 1$, corresponding to the pair of documents having the highest number of links according to the linking method L. The worst weight for $W_L(A, B)$ is the maximum number of outgoing edges from A (bounded by the number of documents in the collection).

The document graphs obtained for each linking


Whole collection with linking method 1
A
5 jumps
B
D
C
2 jumps
Whole collection with linking method 2
A
2 jumps
B
D
C
1 jumps
Documents visited by the user
A
D
User 1
B
C

A → B = 5	C → A = 2
A → C = 5	C → B = 8
A → D = 5	C → D = 2
B → A = 6	D → A = 6
B → C = 6	D → B = 6
B → D = 6	D → C = 1

Average number of jumps = 4.83

A → B = 2	C → A = 2
A → C = 4	C → B = 4
A → D = 4	C → D = 1
B → A = 3	D → A = 1
B → C = 2	D → B = 3
B → D = 2	D → C = 5

Average number of jumps = 2.75

Figure 3: Example of number of jumps (without weight for clarity) between documents in a user log, compared between two linking methods on the collection. Note that because our graph is directed, the number of jumps between two documents, e.g. A and B (5 for method 1) will be different than between B and A (6 for method 1).

method are evaluated by their ability of visiting efficiently, by taking into account the weights previously described, each user-log collection of documents. As we can see in figure 3 where no weight is specified for clarification purpose, to visit all documents selected by user 1, we need to follow on average 4.83 links with the linking method 1 and 2.75 with method 2. Therefore, we will consider that the linking method 2 is more efficient for recreating the journey of user 1 than method 2.

Metric

To evaluate the quality of our different methods on the graph produced, we use the average number of jumps corresponding to the average length in terms of edges in the weighted shortest paths between all node pairs in a set of documents.

This metric can be viewed as the number of links to be clicked on in a source document to reach the target article, or the number of intermediate articles to be visited, counting the initial one, as shown in figure 3. Since weights in our graph are in increasing order of importance (the best weigh is 1), finding the length of the weighted shortest path between documents A and B gives insight into the likelihood of a reader navigating from A to B via recommendation links.

We can then compare these values across methods and juxtapose them with the average number of weighted jumps between all document pairs in the collection.

5. Experiments

5.1. User log collection

Three test sessions were held in May, June and November 2023 to test the *ARCHIVAL* demonstrator with potential users. During the first test day a total of sixteen testers came together for a experiment of discovery and familiarization with the *ARCHIVAL* system. Two panels were set up to carry out two test sessions: the morning session brought together ten testers, mainly researchers in information and communication sciences, while the afternoon session was made up of six testers with a profile of library and documentation professionals. Four testers were invited for the second session of experiment: two information and communication sciences teacher-researchers, a PhD student and teacher-documentalist, and a librarian. Then the third session gathered five expert researcher in the domain of the OCRized journal. For each test session, the general framework was the same and we consider all testers to be part of a single panel of 25 users.

During each test session, testers were instructed to explore the demonstrator freely. After 40 minutes of free exploration where we observed their use of the interface, they answered an initial general questionnaire on their apprehension and appropriation of the device, their use of certain functionalities and their documentation habits. The testers then continued the experiment using suggested entry articles.

5.2. Linking strategies

We performed question generation and linking on the 896 documents used in this study. The average number of questions generated by the BARThez model for each granularity level are given in table 1 as well as the percentage of elements containing at least one question for each level. We can see that about 16% of the documents do not contain any question, this corresponds mainly to the summaries or bibliography where we could not generate questions. Less than half of all paragraphs contain at least one question, with an average of 1.0 questions per paragraph and 2.7 if we exclude paragraph with no questions at all. The 60.4% of paragraph that doesn't contain any question consists either of very short ones such as end notes, titles and all micro-textblocks detected by the OCR or of paragraphs where our question filtering process discarded all the questions generated as being non relevant.

For comparison, a second question generation method based on a larger model, Mistral-7B-Instruct-v0.2 (Jiang et al., 2023) was employed. This time, generation was conducted directly at the page level, without prior extraction of potential answers, utilizing an empirically created prompt (as see in listing 1) and no additional filters. Pages were selected based on the presence of at least one question generated by the BARThez model for generation with Mistral-7B-Instruct-v0.2. We aligned the number of generated questions to the one obtained with the previous approach, a portion of questions were randomly removed to ensure balanced comparison. Subsequently, an average of 71.6 questions were generated per article using this method.

Listing 1: Mistral prompt

```
You're a professor of history in the
     field of human and social
    sciences. Annotate the document
    in the form of open questions in
    French as you read about key
    elements of the given paragraph.
    The questions shouldn't be too
    verbose, and may relate to
    elements whose answers are
    present in the paragraph or not.
{document}
Questions :
-
```

Following the methodology in section 5 We have generated four document linking graphs, as described in figure 3, one for the question-linking method G_{qa} using both questions and answers to compute similarity measures, and one for the paragraph similarity method G_{para}. The two other ones correspond to the graphs produced by the same method, applied only on the questions of BARThez G_q and Mistral $G_{mistral}$, the latter having no answer extraction.

Measure	Article	Page	Paragraph
avg. nb. Q. per element	70.4	9.4	2.7
% elements with Q.	83.8%	84.6%	39.6%

Table 1: Average number of questions generated at each level of granularity (document, page, and paragraph) for the BARThez model and percentage of articles with at least one question

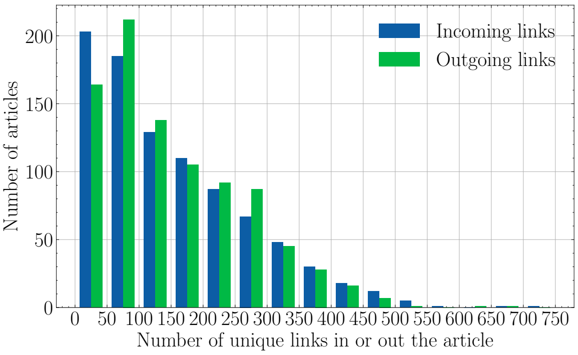


Figure 4: Count of document by number of incoming and outgoing links for G_{qa}

To build G_{qa} we computed the cosine similarity metric between all the SentenceBert (Reimers and Gurevych, 2019) embeddings of the concatenation of the question and answer (question + answer) and kept the top 10 links for each of them. The question where generated on 896 articles (corresponding to nodes) and produced 136,478 unique links in total (corresponding to edges).

For G_{para} we computed the cosine similarity metric using the embedding of the paragraphs having produced at least one question. We generated the links with the same constraints as for the questions + answer, with a maximum of 10 links. We produced 129,550 unique links for this methods.

Finally, for G_q and $G_{mistral}$ we computed the cosine similarity metric directly on the embeddings of the generated questions of both models. We generated the links with the same constraints as for the questions + answer, with a maximum of 10 links. We produced respectively 147,662 unique links for G_q and 153,454 unique links for $G_{mistral}$.

5.3. Results

All our constructed graphs feature a single, strongly connected component. This result shows that it's possible to explore the entire collection using the links produced by all the used methods, without getting stuck in a clique.

In our experiments we have 25 users, so we used 25 user-log collections, with an average of 13.7 documents in each set.

	#links	Average number of jumps	
Graph	avg.	All articles	User logs
G_{para}	144	5.18	4.34
G_{qa}	152	4.88	4.01
G_q	165	4.69	3.83
$G_{mistral}$	174	5.26	4.34

Table 2: Average number of links (in and out) for each method and average number of jumps for all pairs of articles in the entire collection (All articles) and in user-log collections

The average number of unique links in and out of each document is given in table 2. A more precise breakdown of articles according to their number of incoming and outgoing links for G_{qa} is shown in figure 4.

We can see in table 2 that for all methods, the average number of jumps between the users articles is lower than the average number of jumps between articles in the collection. We can assume that those methods gives the user easier access to articles considered relevant than to a random article, the link using question being the one bringing explored articles closer together.

BARThez's question-only linking method gives the best results over the other results, and specifically over his question+answer variant, with the shortest average path. This result is consistent with feedback from platform users who told us that they didn't find the answer useful in their search for links, and that it could even confuse them.

The questions produced by Mistral do not yield links as dense as the other question generation methods, with scores close to the one of the similarity between paragraph. This can be explained by several factors, the first being the granularity of the generation, at page level rather than paragraph level. The second is the generation method and task, with prompting for more open-ended and general questions than SQuAD-style text comprehension questions with already-defined answers. The last is the absence of an expert filter on question generation, as described in section 2.

These experiments show that it is possible to use logs from users exploration in order to compare and evaluate linking methods as an extrinsic task for evaluating the usefulness of question generation methods. The results obtained can give some indications about the efficiency of finding connected documents with a given linking method.

6. Conclusion

In this paper, we introduced a framework and approach for harvesting ecological user logs to evaluate a question generation method trough an extrinsic document linking task. By exploiting graph metrics, we conducted evaluations using these logs to gain insights into the links generated by our method, comparing them with links produced by a LLM and traditional linking techniques. Our results highlight a notable observation: even a compact model such as BARThez, enhanced with expert filters and heuristics, can outperform a generic-purpose LLM in generating task-specific questions. This underscores the effectiveness and robustness of our methodology in enabling a comparison of questions through an extrinsic document linking task, offering insights into the efficacy of various question generation approaches trough this specific task. The data collected in this study is available on the following link: https://gitlab.lis-lab.fr/archival-public/autogestion-qa-linking.

Acknowledgements

This work has been partially funded by the Agence Nationale pour la Recherche (ANR) through the following program: *ARCHIVAL* ANR-19-CE38-0011. This project was provided with computer and storage resources by GENCI at IDRIS thanks to the grant 2023-AD011012688R2 on the supercomputer Jean Zay's V100/A100 partition.

Elie Antoine, Hyun Jung Kang, Ismaël Rousseau, Ghislaine Azémard, Frédéric Bechet, and Géraldine Damnati. 2023. Exploring social sciences archives with explainable document linkage through question generation. In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 141–151.

Frederic Bechet, Elie Antoine, Jérémy Auguste, and Géraldine Damnati. 2022. Question generation and answering for exploring digital humanities collections. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4561–4568, Marseille, France. European Language Resources Association.

Iva Bojic, Jessica Chen, Si Yuan Chang, Qi Chwen Ong, Shafiq Joty, and Josip Car. 2023. Hierarchical evaluation framework: Best practices for human evaluation. *Human Evaluation of NLP Systems*, page 11.

Robin Brochier and Frédéric Béchet. 2021. Predicting links on wikipedia with anchor text information. In *Proceedings of the 44th International*

ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '21, page 1758–1762, New York, NY, USA. Association for Computing Machinery.

Martin d'Hoffschmidt, Wacim Belblidia, Quentin Heinrich, Tom Brendlé, and Maxime Vidal. 2020. FQuAD: French question answering dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1193–1208, Online. Association for Computational Linguistics.

Shasha Guo, Lizi Liao, Cuiping Li, and Tat-Seng Chua. 2024. A survey on neural question generation: Methods, applications, and prospects. *arXiv preprint arXiv:2402.18267*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Moussa Kamal Eddine, Antoine Tixier, and Michalis Vazirgiannis. 2021. BARThez: a skilled pretrained French sequence-to-sequence model. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9369–9390, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Comput. Linguist.*, 31(1):71–106.

Valentina Pyatkin, Paul Roit, Julian Michael, Yoav Goldberg, Reut Tsarfaty, and Ido Dagan. 2021. Asking it all: Generating contextualized questions for any semantic role. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1429–1441, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.