Exploring Social Sciences Archives with Explainable Document Linkage through Question Generation

Elie Antoine, Hyun Jung Kang, Ismaël Rousseau, Ghislaine Azémard, Frederic Bechet, Géraldine Damnati

Exploring Social Sciences Archives with Explainable Document Linkage through Question Generation

Elie Antoine[1], **Hyun Jung Kang**[2], **Ismaël Rousseau**[2], **Ghislaine Azémard**[3]
Frederic Bechet[1], **Géraldine Damnati**[2]
(1) Aix Marseille Univ, CNRS, LIS, France {first.last}@lis-lab.fr
(2) Orange Innovation, DATA&AI, Lannion {first.last}@orange.com
(3) FMSH/Univ Paris 8 Chaire UNESCO ITEN azemard@msh-paris.fr

Abstract

This paper introduces the question answering paradigm as a way to explore digitized archive collections for Social Science studies. Question generation can be used as a way to create explainable links between documents. Question generation for document linking is validated on a new corpus of digitized archive collection of a French Social Science journal.

1 Introduction

From an information and communication science perspective, two steps are essential to bring together computer technology and human and social science objectives. The first essential step is the availability of annotated data in order to train and evaluate with objective metrics the model deployed to ensure their relevance and scientific interest. The second essential step is the creation of an interface adapted to the objectives of the device and respectful of the user by putting forward the explainability of the results provided.

This methodology was followed in the context of the *Archival* project[1]: firstly existing annotated data have been used to train and evaluate deep neural network question generation models; secondly we applied these models to a corpus of social science archives for evaluating their relevance in a real archive exploration application. This paper describes this second step, which study if recent advances in Natural Language Processing thanks to deep learning models translate into novel mediation interfaces for social science researchers.

This paper is structured as follows: section 2 presents our methodology for generating explainable links among documents based on question-generation models; section 3 presents the archive corpus used in this study; section 4 presents our question generation and filtering method; section 5 describes how generated questions can be used to create explainable links within documents; finally sections 6, 7 and 8 presents an experimental study on our archive corpus with quantitative, descriptive and qualitative evaluations of the method proposed.

2 Exploration through questions generation

When exploring a thematic archive collection, links can be made between documents or parts of documents according to various criteria such as co-occurence of entities (person, location, organisation, date, ...), keywords related to a knowledge base or a thesaurus (Tsatsaronis et al., 2014), or directly by a statistical similarity measure between documents or parts of documents such as sentences (Wang et al., 2016) or paragraphs (Dai et al., 2015).

Furthermore, some methods produce links directly between the embedding of the whole documents (Ginzburg et al., 2021), (Jiang et al., 2019), the sum of the word embeddings of the document (Landthaler et al., 2018) or by representing them with word-graphs and using shortest-path algorithms for linking (Nikolentzos et al., 2017). The graph structure obtained can then be used to design navigation interfaces such as maps representing linked documents or directly by inserting hypertext links (Mihalcea and Csomai, 2007; Brochier and Béchet, 2021).

The weaknesses of keywork/entity links are on the one hand the amount of links generated that can be very big if large sets of keywords or entities are considered and on the other hand the fact that the simple occurrence of relevant terms does not mean that their contexts of occurrence are relevant or interesting to users. On the contrary, similarity-based links take words in context into consideration, but the use of statistical similarity metrics make the links often difficult to interpret.

Recently, advances in Question Answering (QA) models from text have enabled the use of asking

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

direct natural language questions in order to access to electronic documents. Impressive results have been obtained with current deep learning language models on benchmark corpora such as SQuAD (Rajpurkar et al., 2016), however it has been shown that the kind of questions that these models handle best are simple literal questions for which a factual answer can be found in the text and that performance drops when dealing with more abstract questions or questions needing a larger context than a sentence in order to be addressed. Moreover, most of these studies have been applied only on Wikipedia text. A recent study (Bechet et al., 2022) have shown that *realistic* questions, like those that can be asked by a professional reader analyzing social science archives are quite far from the simple benchmark questions used to evaluate QA systems, leading to poor performance.

However, even if current QA models might be too simplistic for use in a real archive exploration setting, we believe that Question Generation models can still be useful in order to characterize documents. Such models can be trained on the same corpora as QA models: while QA models are trained to generate a response given a question and a text document, Question Generation models are trained to predict a question given an *answer* and a text document. By selecting potential answers on text segments and generating questions from these answers and their context of occurrence, we obtain an abstraction of a text segment which contains the set of questions that can be asked on it. By estimating similarities between questions and answers belonging to different documents, we can predict links between them that can be explained by the two QA pairs, adding an explainability layer to the process. We believe that this is an efficient way of presenting links to a user: by looking only at the linked QA pairs, readers can decide if it is worth or not to follow this link, saving time compared to the standard solution consisting of following every link to decide if the similarity between two text segments is interesting or not.

In this study, we developed a question generation and filtering process which is used to obtain links between documents of a collection of social science archive corpus. This study presents the first quantitative and qualitative evaluation done of this method on this archive corpus.

3 The *self-management* archive corpus

In order to assess the previously described "exploration through questions generation" paradigm, we have chosen to focus on a particular type of Social Science archive source: a full collection of the *Autogestion* ("self-management") journal published for 20 years between 1966 and 1986. The originality of our work is hence to propose a way to access this rich source through explainable links.

The "self-management" notion falls within the large spectrum of social sciences. It concerns daily social environment, economic life, as well as political life, education, ecology, culture, architecture . . . Nowadays, “self-management” supports in an underlying way the concepts of radical democracy, confederalism, social and solidarity economy and sustainable development. As a source of social innovation, self-management has variations all over the world and questions societal and economic models of development. It is a particularly transversal and interdisciplinary notion which can feed research in sociology, political science, economy, law, political anthropology and social history. The *Autogestion* journal [2] 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[3] foundation’s library. The full collection has been granted the Collex label (*Collection d’Excellence* or Excellency Collection) from the CollEx-Persée[4] network under the supervision of higher education and research for the preservation of corpus of digitized or natively digital documents. (Weill, 1999) describes the journal as an observatory of liberation movements and states that it « *accompanied — preceeding and following — the liberation movement which called for workers’ self-management. Through analysis of its precursors, contemporary practices and historical precedents, the journal was a conceptual tool capable of inspiring action. Its disappearance coincided with the abandonment of the reference to workers’ self- management in socio-political movements, although the aspiration it represented continues to exist.*»

We are using an OCRized version of the corpus. The structure of the journal is rather standard

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

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

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

(mono-column, few figures) and the quality of the OCR provided by Tesseract is sufficient to be exploited as is without manual corrections. Studying the impact of OCR errors is outside the scope of this study and should be investigated in further research work.

The resulting corpus is composed of 46 issues of the journal, ranging over 20 years, for an overall amount of 6298 pages and 1.98M tokens.

4 Question generation

The Question-Generation (QG) task is a classical NLP task that has been revisited thanks to the development of efficient deep learning sequence-to-sequence models (Du et al., 2017; Shakeri et al., 2020; Murakhovs'ka et al., 2022) and Large Language Models (Agrawal et al., 2022). It can be modeled as a neural generation task, where a sequence-to-sequence model is trained to *translate* a sequence of words representing a text segment (the *context*) containing an *answer* (a sub-sequence of words belonging to the *context*) into another sequence of words representing a question on the input. The task is then to generate a question given a *(context, answer)* pair. The availability of large databases of question/answer/context triplets such as SQuAD can be used to directly fine-tune sequence-to-sequence generation models such as *BART*.

One of the key decision that has to be made before generating a question is the choice of the *answers* on which questions will be generated. Choosing all noun phrases as answers can lead to an over generation of questions, most of them being not very relevant if the contexts of occurrence of answers is not informative. That is why we chose to use semantic annotations in order to select answer candidates in order to generate more informative questions. As proposed in (Pyatkin et al., 2021) and (Bechet et al., 2022), we use a Semantic Role Labelling (SRL) model following the PropBank formalism (Palmer et al., 2005) in order to select answers candidates among the semantic roles detected.

In our study, we train the question generation model 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). This is a three steps process:

1. Annotation of the text corpus with Semantic Role Labelling (SRL) labels following the PropBank formalism
2. For each question-answer-context triplet:
   (a) Identification of the semantic role that corresponds to the answer of the given question through the alignment of gold answer spans and semantic role spans, selecting the one with maximum overlap
   (b) Generation of a training example, with the selected answer, current sentence as the context, additional semantic information derived from the semantic role analysis as the input sequence, and the question as the output sequence
3. Fine-tuning of the pre-trained generation model on the collected corpus.

At inference time, generating questions on a given sentence involves performing semantic analysis on the sentence, generating an input sequence for each detected semantic role, and using the fine-tuned seq-to-seq model to generate a question for each input sequence.

The following translated example is from the *FQuAD* training set with ANS being the *answer*, LU the lexical unit that triggers the semantic relation, and CTX the context:

source : `[ANS:ARG2] Héra (Hera)[LU] appelé (called) [CTX] Cérès fut également appelé Héra en Allemagne pendant une brève période. (Cérès was also called Hera in a brief period in Germany.)`

target : `What name did Ceres have for a short time in Germany ?`

The application of the question generation model on a sentence from the *self-management* corpus processed by an SRL parser is illustrated in the following example:

source : `[ANS:ARG0] une bonne partie du C.N.R.S. (a good part of the C.N.R.S.)[LU] évolue (is evolving) [CTX] progressivement une bonne partie du C.N.R.S. évolue vers une structure pour ainsi dire autogérée (gradually a good part of the C.N.R.S. is evolving towards a self-managed structure)`

generation : `Quel organisme évolue vers une structure autogérée ? (Which organization is moving towards a self-managed structure ?)`

We apply a series of filters to enhance the quality and reduce the quantity of generated examples. The first step (**F1**) is to restrict the SRL analysis to

only include frames with a strictly verbal trigger (rejecting auxiliary verbs), as these are deemed to be of higher quality due to their ease of detection.

To further improve the quality of the generated examples, we apply a filter (**F2**) on the queries to remove those with non-informative answers or contexts. This includes answers that are less than 5 characters, or belonging to the NLTK (Bird et al., 2009) stopwords list in order to eliminate answers containing only pronominal coreferences. Queries with a context of fewer than 5 words are also filtered out.

The generated questions are also subjected to a filter (**F3**) based on the "roundtrip consistency" methodology proposed by (Alberti et al., 2019). This filter involves retaining only the synthetic examples where a QA model [5] is able to retrieve a portion of the target answer from the generated question. We consider that the model has successfully retrieved the answer if there is a minimum overlap of 30% between the predicted answer and the answer of the query.

Finally, we apply a final filter (**F4**) to eliminate duplicate questions, which are a frequent phenomenon due to slight variations in some queries, often resulting in very similar or identical questions.

5 Generating explainable links

The main originality of our approach is the use of our synthetic questions/answers to establish links between documents in our corpus. While traditional methods involve computing similarity via document embeddings at a chosen level of granularity (sentence, paragraph or textblock, page), our approach involves computing a similarity measure between question+answer (Q+A) embeddings. We consider a *source* and a *target* Q+A embeddings obtained by the concatenation of questions and answers produced by our method described in Section 4.

For example, this is the "<question> | <answer>" structure obtained on the example of the generated question given in the previous section:

Quel organisme évolue vers une structure autogérée ? (Which organization is moving towards a self-managed structure?) | une bonne partie du C.N.R.S. (a good part of the C.N.R.S.)

[5] in our case, a *CamemBert-large* (Martin et al., 2020) model trained on *FQuAD*

Our embedding projection for each Q+A pair use the SentenceTransformer (Reimers and Gurevych, 2019) library [6]. A cosine similarity measure is then employed between all pairwise combinations of these embeddings, resulting in the computation of a similarity matrix.

For each Q+A pair in our corpus, we extract the 49 most similar pairs across three granularities:

1. The entire collection, including the same page/sentence **(ALL)**
2. All documents within the same issue but in a different article **(OUT_ARTICLE)**
3. All documents outside the current issue **(OUT_NUM)**.

This method allows us to enrich the documents in our archive corpus with many links at different levels with an *explanation* for each link, represented by the two question/answer structures kept after filtering by means of the similarity metric.

In the archive exploration prototype developed for the *Archival* project, these links appear when a user highlight a portion of text in the original document: a window appears with a list of links to other documents from the same archive collection. Each link is *explained* by showing the source and the target questions used to produce the link, as well as a snippet of the target document containing the answer to the target question. The metadata (title, author, date) of the target documents are also displayed. This list of links is sorted according to the similarity metrics between source and target Q+A as well as several heuristics: Q+A containing named entities and terms from a thesaurus attached to the archive collection receive a positive score while Q+A containing coreference mentions receive negative score.

Thesaurus and coreference detection are presented in the next section as well as the analysis of the corpus of questions and links generated, both at a quantitative and qualitative level.

6 Quantitative and qualitative description of the generated questions

In this section, we analyze the application of our generation and filtering method to our *self-management* corpus. We provide first a quantitative study of the set of generated questions followed

[6] we use the multilingual model *distiluse-base-multilingual-cased-v1* (Reimers and Gurevych, 2020)

Filter	F1	F2	F3	F4
Nb. questions	247,907	193,685	129,119	79,869

Table 1: Summary of the different filters: F2 (remove non-informative answers), F3 (round-trip consistency), F4 (remove duplicates).

by a more in-depth descriptive study according to three criteria: question types, question themes and coreference chains.

6.1 Quantitative description

We applied our question generation method on a subset of 24 journal issues of the *Autogestion* collection ranging from 1966 to 1979. Each issue contains several short or long articles, for a total of 448 articles. Since the electronic version of this corpus is obtained through OCR, we have two additional level of segmentation: *page* (corresponding to the OCR of each image of a given page of the collection) and *textblock* (the minimal unit of coherent text output by the OCR system). We consider here a subset of the whole corpus presented in section 3. This subset contains 4786 pages, 33551 textblocks for a total of 1,5M tokens. Initially the Semantic Role Labeling process yields 143,317 Frame detections which is reduced to 124,925 detections when focusing on non-auxiliary verbs as of the **F1** filtering process. Each Frame detection yields an average of 1.7 Frame Elements, meaning that the first set is made of 247,907 questions. Table 1 provides the number of generated questions following the filtering processes described in 4.

As we can see, the total number of questions kept after the four-stage filtering process is 79,869. The average number of questions for each granularity level are given in table 2 as well as the percentage of elements containing at least one question for each level. We can see that about 8% of the articles do not contain any question, this corresponds to the summaries or bibliography where we could not detect Frames and therefore generate questions. More than half the textblocks contain at least one question, with an average of 6.2 questions per textblock. The 48.2% of textblocks that doesn't contain any question consists of very short ones such as end notes, titles and all micro-textblocks detected by the OCR.

6.2 Qualitative description

In this section, we analyze the structure and the content of the automatically generated questions.

measure	article	page	TextBlock
avg. nb. Q. per element	258	25.2	6.2
% elements with Q.	91.7%	95.2%	51.8%

Table 2: Average number of questions generated at each level of granularity (item, page, and textblock) and percentage of items with at least one question

6.2.1 Question types

First, we analyze which type of interrogative pronouns is most often used in the synthetic questions.

We can see in figure 1 that the most used interrogative pronoun is the pronoun "What", with a combined 45% of questions using it. This may be due to several things, the first being that many French words correspond to What (or Which), directly increasing the proportion of this class. The second is based on our training corpus, used twice in our process: in the training of our question generation model and in our roundtrip consistency filtering step. Indeed, in the latter, the proportion of "What" questions is very similar to ours (47.8%). We can assume that our question generation model is biased in this direction and that our filtering method, based on the same dataset and having seen more examples of this type, performs better in the MRQA task on questions of this type, and thus amplify this bias. However, this is also consistent with the distribution of ARG0 and ARG1 arguments predicted by the semantic role labeler.

In second place, a quarter of the generated questions are about a person or a group of persons with the pronoun "Who". This seems consistent with the fact that these types of entities are generally best detected by language models. This may also be related to the fact that our corpus contains many accounts of historical events, of positions taken on various influential characters or movements, thus mechanically increasing the number of questions of this type. To support this possibility, we can see that the *FQuAD* dataset contains only 12.2% of such questions, allowing us to rule out a bias similar to that of the "What" questions.

6.2.2 Question themes

We qualify the themes of the question generated with respect to a specific thesaurus which has been created on the *self-management* domain. Starting from prior knowledge of the domain, a first list of notions has been built. It has then been enriched by a list of keywords and keyphrases extracted from the articles of the *Autogestion* journal. These terms

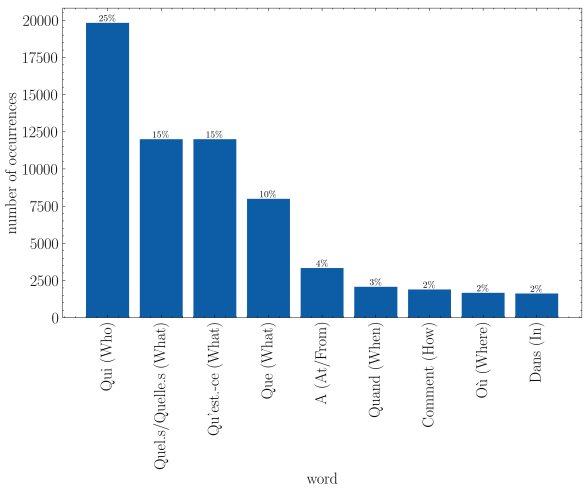


Figure 1: Number and percentage of total number of the interrogative pronoun

$\|w \in T\|$	0	1	2	3	4	5
$\|Q + A\|$	43693	25159	8472	2129	355	58

Table 3: Distribution of the number of words (w) belonging to thesaurus T among questions+answers ($Q + A$)

entry (Fr)	**entry**	**occurrences**
travailleurs	workers	3247
travail	work	2668
pouvoir	authority	2214
société	society	1947
révolution	revolution	1834
production	production	1742
contrôle	control	1470
système	system	1426
ouvrier	laborer	1387
mouvement	movement	1362

Table 4: The ten most frequent thesaurus entries

are mostly nominal phrases extracted thanks to a morphosyntactic analysis of the documents. When flexional variants of a locution are encountered, the form which has the largest number of occurrences is chosen (majority form). From all the extracted keyphrases, the experts have selected a list of additional thesaurus entries, by choosing terms that refer to general notions that can be relevant to index documents. The thesaurus is then sorted hierarchically in order to form a tree structure of maximum of 4 levels depth. The tree has 437 leaves and is organized in 8 general notions at the root of the tree (*Organisations, Social Classes, Economic Development, Exercice of Power, Justice, Political Models, Psycho-sociology, Social Values*).

We analyzed our corpus of synthetic question/answer pairs to investigate the usage of thesaurus entries. Our results show that **30.6%** of the generated questions and **25.1%** of the answers contain at least one term from the thesaurus. Furthermore, we found that **45.3%** of the question-answer pairs included at least one thesaurus word in either the question or the answer, and **10.4%** contained a thesaurus word in both.

A more detailed description of the distribution of the number of entries detected in the questions and answers can be found in Table 3 and the 10 most frequent entries in Table 4.

The pair with the most thesaurus terms is the following :

`Q : Qu'est ce qui rendra possible le` **`développement`** `de la` **`participation`** `des` **`travailleurs`** `et de leurs` **`organisations`** `à la direction et à la gestion des entreprises nationales ? (What will make it possible to` **`develop`** `the` **`participation`** `of` **`workers`** `and their` **`organizations`** `in the direction and management of national enterprises?)`

`A : le` **`changement`** `– en` **`droit`** `et dans les faits – des formes de la` **`propriété`** `(the` **`change`** `- in` **`law`** `and in fact - of the forms of` **`ownership`**`)`

This analysis suggests that apart from allowing the creation of links to explore the collection, generated questions could also be a way to illustrate the main notions that are addressed in the journal. Dedicated interfaces could be developed for this purpose in future work. Additionally, we aim to explore the potential of using those results to filter generated questions and weight links to favor those containing key notions of the corpus as we consider that these questions are more likely to refer to a meaningful concept with respect to the theme of the archive collection explored.

6.2.3 Coreference chains

We are also interested in the impact of coreference chains in our question generation process. Indeed, if a question or an answer contains a sub-specified element of a coreference chains, this can affect the quality of the questions generated and furthermore the relevance of the proposed links. Therefore, we applied a coreference resolution system to our corpus in order to qualify this phenomenon in our set of generated questions.

Modern coreference resolution systems adopt an end-to-end architecture, which integrates mention detection and coreference resolution into a single system. (Lee et al., 2017, 2018) were the first to

propose such an architecture by considering all possible spans of text in the document and assigning coreference links based on the mention score between a pair of spans. There are also end-to-end coreference resolution systems for French, such as DeCOFre (Grobol, 2020) and coFR (Wilkens et al., 2020). DeCOFre[7] is trained primarily on spontaneous spoken language (ANCOR corpus, (Muzerelle et al., 2013)), while coFR[8] is trained on both spoken (ANCOR corpus) and written language (Democrat corpus, (Landragin, 2016)). For this study, we use coFR, as it is better suited for our corpus (i.e., archives and documents).

coFR produced coreference chains, each consisting of a set of mentions that refer to the same discourse entity, for instance: {*la participation au régime capitaliste*, *elle*, *La participation*, *elle*}. For the purpose of the study, we further detect the targets (*i.e.* the most representative mention) in the coreference chains. The longest mention of a coreference chain is chosen as the TARGET of the entity (or the first in case of two equally long mentions). An example for the chain mentioned above is that the TARGET is '*la participation au régime capitaliste*'. In cases where all mentions in a chain are pronouns or determiners (e.g. {*elle*, *son*, *sa*}), then the TARGET is considered to be "NONE". When there are multiple longest mentions of equal length, the first one is selected as the TARGET, e.g. the TARGET for {*Parti communiste*, *PCF*, *PCF*, *du Parti*} will be '*Parti communiste*'.

We analyze here the presence of mentions and targets in the question/answer pairs. Over half of the answers (51.3%) contain a mention, with 12.6% of the responses being entirely a co-reference, as for the questions, 16.6% of them contain mentions. These results suggest that performing co-reference resolution could enhance the similarity calculation and lead to more contextually grounded link suggestions.

7 Qualitative evaluation of the generated questions

In addition to the quantitative and descriptive evaluation of the generated questions on the *self-management* corpus, we performed a first qualitative evaluation on a subset of the collection.

In order to evaluate the quality and relevance of the generated questions, we annotate the generated questions according to two dimensions. The first dimension focuses on the quality of the question form, with questions being categorized as "Valid", "Incoherent question", or "Ungrammatical question". In the second dimension, we assess the relevance of the question after it has been validated in the previous dimension. This evaluation involves three 5-Point Likert scales:

1. "The highlighted segment corresponds well to an answer to the question"

2. "The question is relevant in the context of the sentence"

3. "The question is relevant in the overall context of the reading"

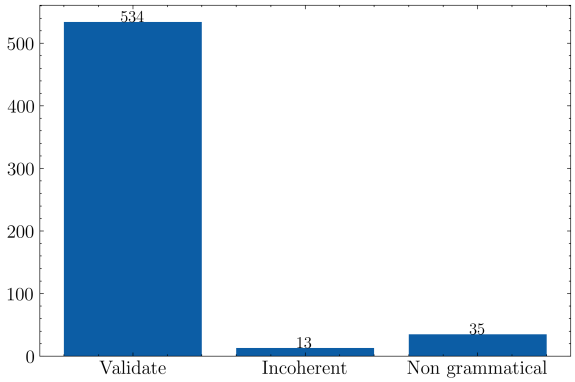


Figure 2: Evaluation of the quality of the form of the generated questions

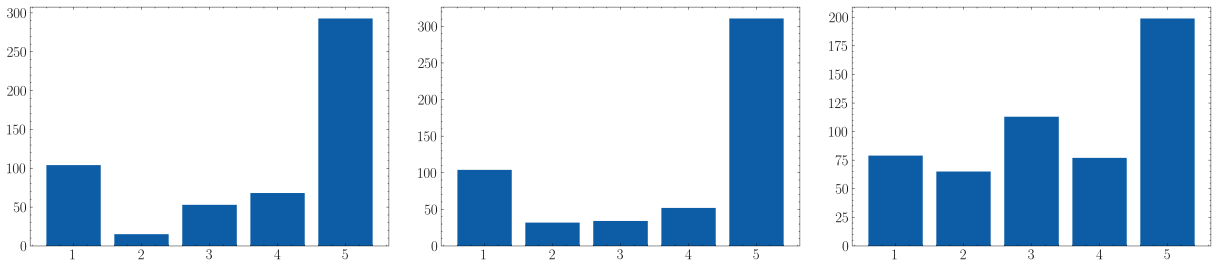
Professional annotators were hired for this task, they annotated a total of 582 questions. As shown in Figure 2, about 92% of the questions were validated on their surface form, which confirms the syntactic quality of our question generation system.

For the relevance annotations, the results are also promising. In terms of answer adequacy (Figure 3(a)), the majority of questions (67%) received a score that indicates a high level of adequacy [9]. The two Likert scales measuring question relevance are more subjective, but a large proportion of questions (over 68%) were rated as relevant in the local context (Figure 3(b)). In the global context (Figure 3(c)) of the reading, the percentage of questions rated as relevant drops, with just over half of the questions meeting the same score criterion.

To check inter-annotator agreement, a subset of 129 questions were annotated by two annotators.

[7] https://github.com/LoicGrobol/decofre
[8] https://github.com/boberle/cofr

[9] In this paragraph, the notion of high level of adequacy corresponds to likert scores > 3



(a) Likert (1) (b) Likert (2) (c) Likert (3)

Figure 3: Evaluation on the relevance of the generated questions

On the first dimension (surface form) we noticed only 11 disagreements between the two annotators. On the second dimension, concerning the first Likert with a simplified 3 category evaluation by grouping the 1 and 2 choices and the 4 and 5 ones, we measured 25 disagreements out of 115 annotations. With the same grouping, we obtain 43 disagreements out of 115 annotations for Likert 2 and 65 out of 115 annotations for Likert 3. These higher numbers of disagreement were expected, as this last evaluation is highly subjective.

8 Document linking evaluation

We quantitatively evaluate the difference in the links generated by the question embeddings compared to two conventional embedding similarity methods. We create a link between two sentences or two paragraphs (textblock) if the similarity between their embeddings is below a threshold (or the top n links are kept).

The comparison results in three "similarity sets":

1. Sentence similarity set **[SENTENCE]**
2. TextBlock similarity set **[TEXTBLOCK]**
3. Question-Answer similarity set **[QA]**

We aim to evaluate the effectiveness of our question-based embedding approach in generating links between documents compared to more traditional embeddings. To quantify the difference, we consider the set of links produced by a question as a unique entity and compute the intersection with the set of links generated by other methods. To ensure that links are compared at the same level of granularity, we consider links identical if they point to the same page. To tackle the fact that one set of links is generated per question for a sentence or paragraph, we aggregate the sets of links for all questions in a sentence or paragraph through a union operation. Finally, the overlap percentage is calculated as the intersection between this union and the corresponding set of links from the sentence or paragraph embeddings.

	Percent of intersection		
Similarity sets	(ALL)	(OUT_ART)	(OUT_NUM)
[QA] // [SENTENCE]	21 %	17 %	19 %
[QA] // [TEXTBLOCK]	23 %	12 %	20 %

Table 5: % of intersection between the similarity sets

This evaluation alone does not allow us to measure the quality of our links. However, it does show (Table 5) that our system produces "original" links through QA embeddings, with nearly 80% of the 49 most similar pages being different from those produced by using similarity methods directly on text segments.

A subjective evaluation which will check the feedback of professional readers to the links and explanation proposed by our method will carried on within the *Archival* project.

9 Conclusion

This paper proposes a new approach for exploring digitized humanities and social sciences collections based on explainable links built from questions. Our experiments show the quality of our automatically generated questions and their relevance in a local context as well as the originality of the links produced by embeddings based on these questions. Analyses have also been performed to understand the types of questions generated on our corpus, and the related uses that can enrich the exploration. Additionally, we discussed the relationships between co-references, generated questions, and extracted answers from the text, which opens a path for future improvements for our system in their resolution. Experiments are still to be conducted to study more qualitatively the generated links, as well as to

enrich and filter in a finer way the large quantity of questions on the corpus.

Limitations

A potential limitation of our method is the use of an SRL semantic framework parser, which can be quite costly to deploy for a very large collection. It would thus be interesting to compare other methods for extracting answers in the test, and for enriching or constraining the question generation.

Our study uses only French monolingual models and corpora, so language does not seem to be a limitation for languages with similar or superior resources.

Additionaly, our study should be pursued to further assess the relevance of links, which necessitates a dedicated evaluation protocole. However we believe that assessing in the first place the quality of generated questions is important for the rest of our work.

Acknowledgements

We would like to thank the reviewers for their insightful comments and suggestions. This work has been partially funded by the Agence Nationale pour la Recherche (ANR) through the following program: *ARCHIVAL* ANR-19-CE38-0011. This work was performed using HPC resources from GENCI–IDRIS (Grant 2022-AD011012688R1).

References

Priyanka Agrawal, Chris Alberti, Fantine Huot, Joshua Maynez, Ji Ma, Sebastian Ruder, Kuzman Ganchev, Dipanjan Das, and Mirella Lapata. 2022. Qameleon: Multilingual qa with only 5 examples.

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. Synthetic QA corpora generation with roundtrip consistency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6168–6173, Florence, Italy. Association for Computational Linguistics.

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.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

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.

Andrew M. Dai, Christopher Olah, and Quoc V. Le. 2015. Document embedding with paragraph vectors. *CoRR*, abs/1507.07998.

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.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1342–1352, Vancouver, Canada. Association for Computational Linguistics.

Dvir Ginzburg, Itzik Malkiel, Oren Barkan, Avi Caciularu, and Noam Koenigstein. 2021. Self-supervised document similarity ranking via contextualized language models and hierarchical inference. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3088–3098, Online. Association for Computational Linguistics.

Loïc Grobol. 2020. *Coreference resolution for spoken French*. Theses, Université Sorbonne Nouvelle - Paris 3.

Jyun-Yu Jiang, Mingyang Zhang, Cheng Li, Michael Bendersky, Nadav Golbandi, and Marc Najork. 2019. Semantic text matching for long-form documents. In *The World Wide Web Conference*, WWW '19, page 795–806, New York, NY, USA. Association for Computing Machinery.

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.

Frédéric Landragin. 2016. Description, modélisation et détection automatique des chaînes de référence (DEMOCRAT). *Bulletin de l'Association Française pour l'Intelligence Artificielle*, (92):11–15.

Jörg Landthaler, Ingo Glaser, and Florian Matthes. 2018. Towards explainable semantic text matching. In *Legal Knowledge and Information Systems*, pages 200–204. IOS Press.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692, New Orleans, Louisiana. Association for Computational Linguistics.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Rada Mihalcea and Andras Csomai. 2007. Wikify! linking documents to encyclopedic knowledge. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, page 233–242, New York, NY, USA. Association for Computing Machinery.

Lidiya Murakhovs'ka, Chien-Sheng Wu, Philippe Laban, Tong Niu, Wenhao Liu, and Caiming Xiong. 2022. MixQG: Neural question generation with mixed answer types. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1486–1497, Seattle, United States. Association for Computational Linguistics.

Judith Muzerelle, Anaïs Lefeuvre, Jean-Yves Antoine, Emmanuel Schang, Denis Maurel, Jeanne Villaneau, and Iris Eshkol. 2013. ANCOR, premier corpus de français parlé d'envergure annoté en coréférence et distribué librement. In *TALN'2013, 20e conférence sur le Traitement Automatique des Langues Naturelles*, pages 555–563, Les Sable d'Olonne, France.

Giannis Nikolentzos, Polykarpos Meladianos, François Rousseau, Yannis Stavrakas, and Michalis Vazirgiannis. 2017. Shortest-path graph kernels for document similarity. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1890–1900, Copenhagen, Denmark. 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.

Siamak Shakeri, Cicero Nogueira dos Santos, Henghui Zhu, Patrick Ng, Feng Nan, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2020. End-to-end synthetic data generation for domain adaptation of question answering systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5445–5460, Online. Association for Computational Linguistics.

George Tsatsaronis, Iraklis Varlamis, and Michalis Vazirgiannis. 2014. Text relatedness based on a word thesaurus. *CoRR*, abs/1401.5699.

Zhiguo Wang, Haitao Mi, and Abraham Ittycheriah. 2016. Sentence similarity learning by lexical decomposition and composition. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1340–1349, Osaka, Japan. The COLING 2016 Organizing Committee.

Claudie Weill. 1999. La revue autogestion comme observatoire des mouvements d'émancipation. *L'Homme et la société*, 132(2):29–36. Included in a thematic issue : Figures de l' « auto-émancipation » sociale.

Rodrigo Wilkens, Bruno Oberle, Frédéric Landragin, and Amalia Todirascu. 2020. French coreference for spoken and written language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 80–89, Marseille, France. European Language Resources Association.

A Examples of generated questions

Examples of generated question kept after the filtering process:

Q : Quel organisme évolue vers une structure autogérée ? (*Which organization is moving towards a self-managed structure ?*)
A : une bonne partie du C.N.R.S. (*a good part of the C.N.R.S.*)
CTX : Mais je crois que progressivement une bonne partie du C.N.R.S. évolue vers une structure pour ainsi dire autogérée, bien que le terme, à ma connaissance, soit rarement avancé. (*But I believe that gradually a good part of the C.N.R.S. is evolving towards a self-managed structure, so to speak, although the term, to my knowledge, is rarely used.*)

Q : Quelle était la conséquence de l'autogestion ? (*What was the consequence of self-management ?*)
A : l'autogestion, si elle limitait leurs droits théoriques en transférant la gestion de l'entreprise à l'assemblée des travailleurs, ne leur offrait pas moins davantage de droits réels de direction de l'entreprise qu'ils n'en avaient jamais eu jusqu'à présent (*self-management, although it limited their theoretical rights by transferring the management of the enterprise to the workers' assembly, did not give them any less real rights to direct the enterprise than they had ever had before.*)
CTX : Les « managers » s'aperçurent que l'autogestion, si elle limitait leurs droits théoriques en transférant la gestion de l'entreprise à l'assemblée des travailleurs, ne leur offrait pas moins davantage de droits réels de direction de l'entreprise qu'ils n'en avaient jamais eu jusqu'à présent. (*The "managers" realized that self-management, although it limited their theoretical rights by transferring the management of the enterprise to the workers' assembly, did not give them any less real rights to direct the enterprise than they had ever had before.*)

Example of queston filtered by (F3) :
Q : Quelle est la nationalité de la Yougoslavie ? (*What is the nationality of Yugoslavia?*)
A : la Yougoslavie (*Yugoslavia*)
CTX : Yougoslavie ait été introduite et se développe pour des causes qui ne seraient pas purement économiques .... (*Yugoslavia was introduced and is developing for reasons that are not purely economic ...*)

B Examples of explainable links

Q_1 : Quel progrès a été réalisé dans l'agriculture avec un capital relativement faible ? (*What progress has been made in agriculture with relatively little capital?*)
A_1 : un progrès très rapide dans la technique et la technologie (*a very rapid progress in technique and technology*)
Q_2 : Qu'est ce qui permet d'augmenter la production agricole ? (*What makes it possible to increase agricultural production?*)
A_2 : méthodes agronomiques modernes (*modern agronomic methods*)

Q_1 : Quel est le but des questions administratives incompréhensibles ? (*What is the purpose of incomprehensible administrative questions?*)
A_1 : à créer leur dépendance (*to create their dependence*)
Q_2 : Quel est le but de la bureaucratie ? (*What is the purpose of bureaucracy?*)
A_2 : ses prétentions à la domination sociale (*its claims to social dominance*)

Q_1 : Qu'est ce qui permet à l'ouvrier gestionnaire d'augmenter son revenu personnel ? (*What allows workers-managers to increase their personal income?*)
A_1 : la productivité du travail (*labour productivity*)
Q_2 : Qu'est ce qui pousse les travailleurs à augmenter la productivité ? (*What drives workers to increase productivity?*)
A_2 : le revenu (*income*)

Q_1 : Quelles entreprises sont en train de se transformer en coopératives de production ? (*Which companies are transforming into production cooperatives?*)
A_1 : les entreprises autogérées et celles qui sont en train de le devenir (*self-managed companies and those in the process of becoming self-managed*)
Q_2 : Quelles entreprises ont eu tendance à perdre leur caractère de coopérative ? (*Which companies have tended to lose their cooperative character?*)
A_2 : Les coopératives ouvrières du XXème siècle (*Workers' cooperatives of the 20th century*)