Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4561–4568
Marseille, 20-25 June 2022

Question Generation and Answering for exploring Digital Humanities collections

Frédéric Béchet[1] Elie Antoine[1] Jeremy Auguste[1] Géraldine Damnati[2]
(1) Aix-Marseille Université, CNRS, LIS {first.last}@lis-lab.fr
(2) Orange Innovation, DATA&AI, Lannion {first.last}@orange.com

Abstract

This paper introduces the question answering paradigm as a way to explore digitized archive collections for Social Science studies. In particular, we are interested in evaluating largely studied question generation and question answering approaches on a new type of documents, as a step forward beyond traditional benchmark evaluations. Question generation can be used as a way to provide enhanced training material for Machine Reading Question Answering algorithms but also has its own purpose in this paradigm, where relevant questions can be used as a way to create explainable links between documents. To this end, generating large amounts of question is not the only motivation, but we need to include qualitative and semantic control to the generation process. In the framework the French ANR project ARCHIVAL, we propose a new approach for question generation, relying on a BART Transformer based generative model, for which input data are enriched by semantic constraints. Question generation and answering are evaluated on several French corpora, and the whole approach is validated on a new corpus of digitized archive collection of a French Social Science journal.

Keywords: Question Generation, Machine reading Question Answering, Digital Humanities

1. Introduction

Recent advances in representation learning of text have achieved remarkable results on benchmark Natural Language Understanding (NLU) tasks as shown in the recent *General Language Understanding Evaluation* (GLUE) benchmarks (Wang et al., 2018), reaching even so-called *human performance* on several tasks (Wang et al., 2019) such as linguistic acceptability, question answering or semantic similarity. However these impressive results are obtained on corpora specifically prepared for these benchmark evaluations; moreover these *understanding* tasks, although always related to a linguistic competency, can be considered as rather artificial as they are tailored to fit the need of system benchmark evaluation and can be quite far from a real application.

In order to study how the current boost in performance in NLU models on benchmark data translates to real-life settings, the applicative framework considered here is the exploration of digitized collections by professional users that are used to analyze archives in order to perform Social Science research. We chose to focus on the question/answering paradigm, as asking questions and looking for answers is at the same time a natural way for researchers to explore archives and also the task that received the most attention in recent language understanding studies, especially since the release of large training data such as SQuAD (Rajpurkar et al., 2016) [1].

In this paper we will present first the *self-management* corpus, a collection of a French journal ranging over 20 years from 1966 to 1986, which has been chosen as our archival material in the ARCHIVAL project, then we will highlight the differences between benchmark corpora usually based on *Wikipedia* and digitized archive collections. We will then present the question/answering paradigm, the annotation scheme developed in Archival and point out the differences between the kinds of questions that can be made by professional users and those used in Machine Reading datasets such as *SQuAD*. We will describe the question generation and question answering models that have been developed to adapt a Machine Reading model trained on Wikipedia to the *self-management* corpus of the ARCHIVAL project without any supervision. Finally we will present the first results obtained on the *self-management* dataset with this adapted Machine Reading model.

2. The self-management corpus

2.1. Origin of the collection

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, ... It addresses populations structure, the relationship of populations with resources, the political, legal and administrative framework of society and the authority relations between individuals and groups. Since the 1960's, the FMSH[2] foundation's library has gathered a pluridisciplinary multilingual mixed collection (archives and documents) about self-management (*autogestion* in French). It gathers around 25000 pieces: books, journals, reports, leaflets, correspondences.

[1] https://rajpurkar.github.io/SQuAD-explorer/

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

2.2. Corpus description

For this study, we are particularly interested in the *Autogestion* journal [3] which is distributed in its digitized form by the French Persée organization. We are using a version of the corpus that has been OCRized with Tesseract without manual corrections. Hence data are not free of OCR errors but the structure of the journal (mono-column, few figures) implies that the OCR quality is good (further studies could imply precise evaluation of OCR quality and impact of OCR errors on downstream NLP tasks but for this study, OCR output are taken *as is*).

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

2.3. Specificities of texts from an NLP point of view

Most studies in Information Extraction or Question Answering are carried out on Wikipedia pages. Wikipedia documents are particularly well suited for theses tasks as they intrinsically dedicated to convey factual information. Another characteristic of Wikipedia is that articles are supposed to follow a Neutral Point of View policy [4]. Recent work (Bertsch and Bethard, 2021) aims at detecting so-called puffery (*i.e* sentences that do not respect that policy, which are tagged by editors as ”peacock phrases”) but this phenomenon remains very rare. On the contrary, texts that are relevant for Digital Humanities and studies related to Social Science are not only factual and neutral documents but also essays or articles that reflect the writer's point of view. Description of events are not only depicted by facts but with deeper analysis of the previous notions or influences that yielded this event as well as their consequences and how they influenced the thinking of other actors.

The following figures provide a few insights of the differences between language in Wikipedia pages and language in the ”Autogestion” journal. Of course a more comprehensive study would be necessary to characterize precisely ”Autogestion” journal texts, but we propose these figures as we believe they can be relevant for our tasks of Information Extraction and Question Answering. Distributions from Wikipedia were extracted from two portals (Archeology and First World War) as gathered in the public French corpus CALOR (Marzinotto et al., 2018) designed for Semantic Frame analysis and Machine Reading Question Answering (Béchet et al., 2019). Figure 1 and figure 2 respectively show the distribution of Part of Speech tags and morphological features, as obtained by the Spacy tokenizer and syntactic parser.

The main differences in terms of POS distribution is that the Autogestion corpus contains less Proper Nouns (PROPN) and less prepositions (ADP). These two observations suggest that sentences are less descriptive. On the other hand it contains more pronouns (PRON) suggesting longer sentences with more anaphoras and more adverbs (ADV) and adjectives (ADJ). The higher proportion of symbols (SYM) however is probably an artefact of OCR errors. The morphological features distributions reveal a higher proportion of plural and feminin forms and a higher proportion of present tense and less past tense.

3. The question answering paradigm

With recent advances in Machine Reading Comprehension (or MRQA for Machine Reading Question Answering) and Knowledge-Base Question Answering (KBQA) along with a very active community participating to competitions and challenges on several public QA benchmark corpora, it is now accessible to use such models in realistic use-cases. In our vision, questioning documents can be a voluntary process performed by users who express a given question on purpose or can be an *implicit* process that helps creating links between documents, transforming a collection of documents into a graph of documents. The former is already implementable through efficient Question Answering Search platforms such as Haystack [5] but the latter is more original and implies that the system can automatically infer relevant questions on documents.

The Question Answering Paradigm can be summarized as follows:

- Questions to a general (or specifically designed) Knowledge Base: Q&A as an assistant for the user to improve its prior knowledge or to facilitate its comprehension of the mentioned notions
- Questions as an advanced search engine, beyond keyword search for the user: finding all documents that might contain an answer to a specific question with the candidate answer highlighted (Text Retriever + Machine Reading Question Answering)
- Questions as an implicit yet explainable generator of links towards related documents: if two documents raise the same questions, or if a document contains the answer to a question raised in another document, there is a link between these documents (these links can be explained by an ananlysis of the questions used to create them).

In the framework of the French ANR project ARCHIVAL [6] we will focus on the last two tasks of this paradigm, therefore we need models that can retrieve answers from users questions and which can generate questions from a text or a segment of text. These models are described in the following section, we will see

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

[4] en.wikipedia.org/w/index.php?title=Wikipedia:Neutral_point_of_view

[5] https://haystack.deepset.ai

[6] https://anr.fr/Project-ANR-19-CE38-0011

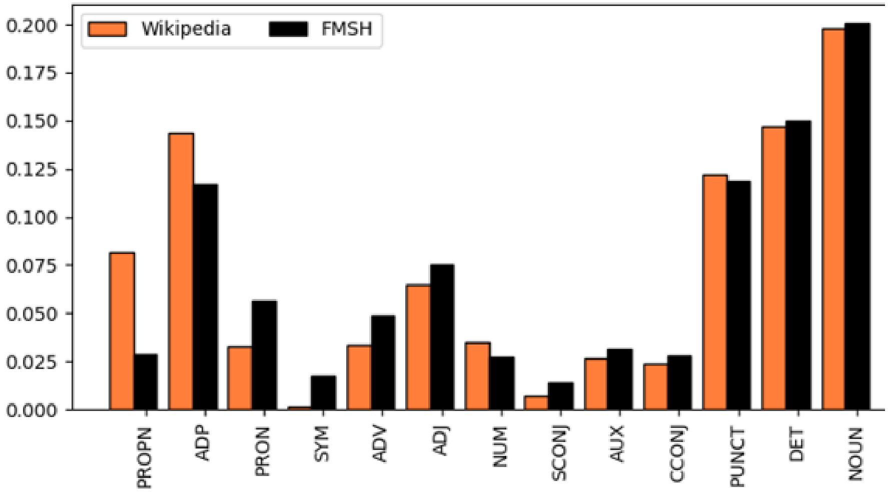

Figure 1: Comparison of Wikipedia and Self-Management (FMSH) corpora according to POS distributions

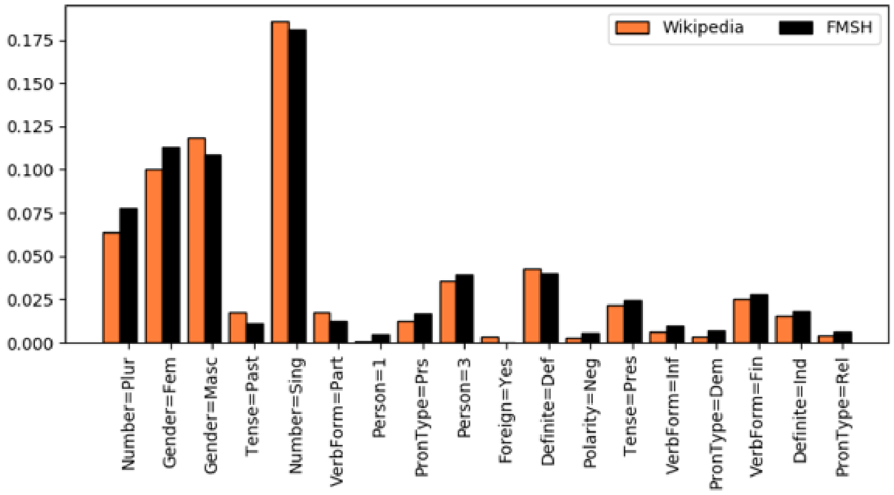

Figure 2: Comparison of Wikipedia and Self-Management (FMSH) corpora according to morphological features distributions

how the question generation model can be used to generate training data for the question answering model.

4. Question generation and answering models

Question generation and question answering are two classical NLP tasks which have been completely rethought with the development of large pre-trained language models. Thanks to generation models such as *BART* (Lewis et al., 2020) or classification models such as *BERT* (Devlin et al., 2018), these two tasks which were traditionally handled through complex linguistic pipelines in the pre-deep-neural-network era have been replaced with straightforward end-to-end approaches with a large boost in performance. In these approaches, the pre-trained models are fine-tuned on the final task thanks to an end-to-end process where the adaptation to the task is done through the choice of the format and the content of the input and output sequences of symbols which will encode the data.

As presented in (Du et al., 2017), the question generation task can be modeled as a neural generation task where a sequence-to-sequence model is trained to *translate* a sequence of words representing a sentence or a passage into another sequence of words representing a question on the input. The task is then to generate a question given a (passage, answer) pair. Large sequence-to-sequence generation models such as *BART* in conjunction with large databases of question/answer/context triplets such as SQuAD can be used to directly train a *passage-to-question* translation model. Training such generation models with generalization capacities beyond existing available reference corpora remains challenging though. (Lyu et al., 2021) propose an unsupervised way of generating synthetic training material for question generation, by using simplified summaries of documents along with simple heuristics to generate domain related training examples of questions that are used to adapt the generation models.

Several approaches have been proposed to make use of synthetic question/answer/context triplets to train Machine Reading models in a data augmentation perspective or in *few-shot* or *zero-shot* settings. In (Béchet et al., 2017), we proposed to use Semantic Frame parsing along with generic patterns in order to generate questions, whose answers would be selected from Frame Elements. (Puri et al., 2020) introduced generative approaches with pre-trained language models by selecting candidate answers with a BERT detection model and generating the corresponding question with a GPT-2 generation model. In a more systematic approach, (Shakeri et al., 2020) extended their approach, predicting (question, answer) pairs from a passage by systematically considering any token as a potential answer. A filtering process based on predictions likelihood is used to select the most relevant questions.

The use of such synthetic question/answer/context triplets have shown to yield improvements in MRQA benchmarks or in adaptation configurations. However the quality of the generated questions remains a drawback. In this work were are not only interested in improving our MRQA model on our Social Science journal corpus, but we are also interested in generating questions that can be used in our question answering paradigm to explore these archival collections. For instance, if we are to propose a link between two passages based on a question they would both provide an answer to, we want this question to be relevant, sounded and correct from a semantic and syntactic point of view. Furthermore we want to be able to explain why this question has been chosen. In this perspective, we will present in this section how we propose to encode the question generation and answering tasks in order to fit this end-to-end paradigm with pretrained models based on Transformer Language Model architectures, while conciliating both qualitative and quantitative objectives for synthetic question generation.

4.1. Question generation model

In our study, following previous work done on the CALOR-QUEST corpus, a semantic representation is added to the sentence as input to the generation model. The goal of this semantic representation is to guide the question generation by explicitly modeling the semantic link between the answers and the arguments of the questions. Two kinds of semantic representation have been tested in this study: a Berkeley FrameNet representation (Baker et al., 1998), following previous work done on question generation on the CALOR-QUEST corpus (Béchet et al., 2019), and Semantic Role Labelling (SRL) following the PropBank formalism as it was proposed to control question generation with *BART* in (Pyatkin et al., 2021).

Training the question generation model from *BART* on a corpus of question/answer/context triplets such as SQuAD is done with the following steps in our study:

1. Annotate with FrameNet and SRL labels the text corpus
2. For each question/answer/context:
   (a) Find the semantic role that corresponds to the answer of a given question thanks to the annotation performed. To do so, we align gold answer spans and semantic role spans and chose the one with the maximal overlap.
   (b) Generate a training example with an input sequence containing the selected answer, the context and eventually additional semantic information derived from the semantic role analysis. The question is the output sequence.
3. Fine-tune the pre-trained generation model on the corpus collected.

At inference time, generating questions on a given sentence consists in first performing semantic analysis on the sentence, then generating an input sequence for each semantic role detected. The fine-tuned seq-to-seq model then generates a question for each of them.

In this study we compare 4 different representations for the input format of the question-generation seq-to-seq model:

1. **basic-Frame-ctx** : the answer is extracted thanks to the alignment process with the Frame Elements as described in step 2.(a) above. The context is simply given as the original sentence with no additional semantic information;
2. **basic-SRL-ctx** : the answer is extracted thanks to the alignment process with the Semantic Roles as described in step 2.(a) above. The context is simply given as the original sentence with no additional semantic information;
3. **full-Frame-ctx**: the answer is extracted thanks to the alignment process with the Frame Elements as described in step 2.(a) above and is further enriched with the *FrameNet* Frame Element label. The context is also explicitly enriched with an extraction of the other Frame Elements and the label of the trigger (Lexical Unit); in this representation we use the Frame Elements labels of the FrameNet lexicon instead of generic roles to characterize the answer and the context.
4. **full-SRL-ctx**: the answer is extracted thanks to the alignment process with the Semantic Roles as described in step 2.(a) above and is further enriched with the *PropBank* Semantic Role label. The context is also explicitly enriched with an extraction of the other Semantic Roles and the label of the trigger (Lexical Unit);

The following example[7] illustrates these 4 representations. In this case, Frame and SRL provided the same span for the answer extraction and the two first configurations are identical.

Context: `Paleolithic tools with teeth and mammoth bones found at [Flins-sur-Seine]answer`

Question: `Where were found paleolithic tools?`

basic-Frame-ctx: `[ANS] Flins-sur-Seine [CTX] Paleolithic tools with teeth and mammoth bones found at Flins-sur-Seine`

basic-SRL-ctx: `[ANS] Flins-sur-Seine [CTX] Paleolithic tools with teeth and mammoth bones found at Flins-sur-Seine`

full-Frame-ctx: `[ANS:Location] Flins-sur-Seine [LU:Locating] found [Sought-entity] Paleolithic tools [CTX] Paleolithic tools with teeth and mammoth bones found at Flins-sur-Seine`

full-SRL-ctx: `[ANS:ARGM-LOC] Flins-sur-Seine [LU] found [ARG1] paleolithic tools [CTX] Paleolithic tools with teeth and mammoth bones found at Flins-sur-Seine`

4.2. Question Answering model

The question answering task on text (Machine Reading Question Answering) consists in detecting the answer to a given question in a text. State-of-the-art approaches consists in fine-tuning a large pre-trained language model such as BERT into the task of predicting the start and end offsets of an answer in a paragraph. The paragraph and the question are given as input features.

To this purpose we used a Machine Reading Comprehension model based on a large language model for French called CamemBERT (Martin et al., 2020), fine-tuned on the question/answering task on different dataset as it will be presented in the experiment section. The question generation model will be used to produce, in an unsupervised way, the training corpus necessary to fine-tune the model to the MRQA task.

5. Corpus annotation

We performed two annotation processes on the self-management corpus. The first one was conducted by professional readers, researchers in social science, that were asked to perform a text analysis on several articles of the *Autogestion* journal. We asked annotators to select areas in the text that they consider as *areas of interest* and put comments that would explain why these areas would be selected. These comments could be key-words, concept or topics related to their analysis or they could be *questions* that were raised by the selected areas.

The second annotation process was conducted by asking annotators specialized in linguistic annotation for NLP projects to collect a question/answer corpus on the same documents. Annotators selected areas in the documents corresponding to answers, and they were asked to write a question for each text area selected. This process is rather similar to the one used through crowdsourcing to build the SQuAD corpus. One significant difference is that we asked annotator to give a *difficulty rating* to each question, 1 being *easy*, very litteral questions (similar to SQuAD), 2 being *difficult* questions were lexical choices were different between the question and the context of the answer, and 3 being *very difficult* questions requiring some abstraction between the text, the question and the segment of text corresponding to the answer. On the overall, 1102 questions were produced but for this study we focused on direct questions and discarded multi-hop questions, resulting in 842 questions.

Our motivation in this double annotation process was that the questions collected through the professional readers are the *realistic* questions and the ultimate goal of a language understanding system, and on the opposite the SQuAD-like collected questions are those that could be handled by current machine reading systems, but that might be too simple or artificial for being of any utility in a real deployed application. By studying the differences between these 2 sets of questions and by measuring how current question/answering systems are affected by changes in lexical choices or abstraction, we hope to open the path to the design of more realistic settings for evaluating language understanding systems.

Annotation	pages	# areas	# quest.
Professional readers	420	2003	1257

Table 1: Annotation performed by professional readers on the self-management corpus

Annotation	# quest.	1	2	3
NLP annotators	842	416	352	74

Table 2: Annotations performed by NLP annotators with a linguistic background.

6. Experiments

All the experiments reported in this study have been made on three corpora:

- FQuAD (d'Hoffschmidt et al., 2020) is a dataset built with the same methodology as the SQuAD

[7] from CALOR-QUEST , translation into English of the French sentence extracted from an illustration caption "*Outils du Paléolithique, avec dents et ossements de mammouth, trouvés à Flins-sur-Seine*"

corpus (Rajpurkar et al., 2016). The 1.0 version used in this study contains 145 articles randomly sampled from a dataset of 1,769 high-quality French Wikipedia articles. A set of 26,108 question/answer pairs have been collected through crowdsourcing. As for SQuAD , this corpus contains mostly *easy* questions with similar lexical choices for the questions and the answers.

- CALOR-QUEST (Béchet et al., 2019) is a question/answer corpus built on top of the CALOR-FRAME (Marzinotto et al., 2018) corpus which contains French encyclopaedic documents (Wikipedia, Vikidia, ClioTexte) semantically annotated following a manual FrameNet semantic analysis of the documents. The test partition of CALOR has been annotated with natural questions with the constraint that the answer matches a Frame Element. This constraint allows us to evaluate MRQA in a semantically controled experimental framework. The test corpus contains 2069 (paragraph, question, answer) triplets manually written by several expert annotators.

- Archival is the self-management corpus presented in the previous section.

In these experiments FQuAD represent the large *generic* corpus that is used to train question generation and question answering models for French. It contains relatively *easy* data (high-quality Wikipedia pages and literal questions) but has the main advantage of being manually labeled and rather large.

CALOR-QUEST is the upper-bound of what we could obtain in terms of MRQA performance with our unsupervised question answering model adaptation process because we use *gold* semantic annotations to generate triplets (paragraph, question, answer) on which a MRQA model can be trained.

Archival is the target corpus with challenging data (OCR noise, difficult subjects, automatic semantic annotation) and more realistic questions.

6.1. Experiments on question generation

The first experiments consists in training a question generation model as presented in section 4.1 on the FQuAD corpus. To this purpose this corpus has been automatically semantically annotated with FrameNet and ProbBank annotations, then 3 different question generation training corpora have been produced: *basic-ctx*, *Frame-ctx* and *SRL-ctx* as described in section 4.1. A seq2seq question generation model based on the BARThez (Eddine et al., 2020) pretrained model is then trained on each corpus.

Table 3 presents the results obtained on the test partition of the CALOR-QUEST corpus in terms of *BLEU* and *BERTScore* (Zhang* et al., 2020) scores between the questions generated by the BARThez model and those manually written on the CALOR-QUEST corpus. We used the default values of sacreBLEU's[8] (Post, 2018) that correspond to BLEU-4 and the configuration of *BERTScore*[9] with a baseline rescaling in all our experiments. As we can see, adding semantic annotation during the generation process seems to improve both the *BLEU* and *BERTScore* scores. However *BLEU* focuses only on the surface forms of the questions and does not reflect their semantics and although BERTScore uses contextual embeddings, it is still based on a single reference question in our case. Human evaluations will be needed in the future to confirm these tendancy on subjective evaluations.

	BLEU	**BERTScore**		
		P	**R**	**F1**
basic-Frame-ctx	20.4	52.2	49.4	50.6
basic-SRL-ctx	21.5	52.4	49.3	50.6
full-Frame-ctx	22.0	54.2	51.6	52.7
full-SRL-ctx	22.3	54.2	51.0	52.4

Table 3: Question generation evaluation with the *BLEU* and *BERTScore* metrics on the test partition of the CALOR-QUEST corpus

The evaluation in terms of MRQA performance of models trained on the generated questions can also give some insights on the relevance of these generation methods. To this purpose we applied this question generation process to the Archival corpus in order to generate a large corpus of context/question/answer triplets on which an MRQA model can be learned. The manual inspection of the questions produced shows that their linguistic quality is very high. However some semantic incoherence can occur, probably when the automatic semantic annotation failed, as we can see in table 4 for example 7.

6.2. Experiments on Machine Reading Comprehension

We carried out experiments on the CALOR-QUEST and Archival corpora with MRQA models trained on the question/answering annotations automatically generated as described in section 4.2.

Table 5 presents MRQA results on the CALOR-QUEST test partition with models trained on generated questions with 3 different models representing 3 different input sequence format. As we can see, unlike BLEU scores, adding semantic annotations during the question generation process doesn't seem to improve MRQA performances. On the overall the approach, with a model entirely trained on automatically generated questions, yields good performances. Even if for CALOR-QUEST the semantic analysis is performed manually, this validates the question generation

[8] sacreBLEU's hash signature of our experiments : ”nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.0.0”

[9] BERTScore's hash signature : ”bert-base-multilingual-cased_L9_no-idf_version=0.3.11(hug_trans=4.5.0)-rescaled”

1	`When did the first Russian soviets take shape?`
2	`What event took place in Brussels?`
3	`Under which government were the first workers' councils formed?`
4	`Who is the founder of the Russian trade unions?`
5	`What is the essential characteristic of revolutionary syndicalism?`
6	`What was Georges Gurvitch's profession?`
7	**`*What nationality is France from?*`**
8	`What was Georges Gurvitch's job?`

Table 4: Examples of generated questions on the ARCHIVAL corpus (translated from French to English)

input-seq.	*Exact Match*	*F1*
basic-Frame-ctx	67.6 (±0.9)	78.1 (±1.5)
basic-SRL-ctx	68.9 (±0.4)	79.2 (±0.6)
full-Frame-ctx	68.5 (±0.4)	78.4 (±0.4)
full-SRL-ctx	66.3 (±1.3)	77.3 (±0.8)

Table 5: Question answering results on the CALOR-QUEST test partition with MRQA models trained on the automatic question generated in 4 different conditions (basic-Frame-ctx, basic-SRL-ctx, Full-Frame-ctx, full-SRL-ctx)

approach as a useful tool to provide MRQA training corpora.

Train	*easy (1)*		*difficult (2)*		*very difficult (3)*	
metrics	EM	F1	EM	F1	EM	F1
FQUAD	37.5 (±0.7)	62.9 (±1.0)	25.0 (±0.3)	54.8 (±0.5)	18.0 (±0.6	48.2 (±3.7)
ARCHIVAL	21.0 (±0.8)	39.9 (±0.2)	9.7 (±0.3)	27.7 (±1.9)	1.3 (±1.1)	20.7 (±1.7)
both	40.6 (±0.9)	63.4 (±0.4)	27.4 (±1.6)	53.8 (±0.7)	19.4 (±0.6)	44.3 (±1.7)

Table 6: Question answering results on the ARCHIVAL test corpus according to the level of difficulty of the questions, trained on FQUAD only, generated questions only (ARCHIVAL), or a combination of both

Table 6 presents MRQA performance on the challenging ARCHIVAL test partitions according to the training corpus used for the question answering model. Here, contrarily to the previous table, the semantic analysis used to extract potential answers and to guide the generation process are produced with an automatic semantic parser. As expected, MRQA performances degrade as the level of difficulty of questions increases. We can also see that relying only on generated questions (**ARCHIVAL**) is much worse that using a large *off-the-shelf* generic corpus such as FQUAD . But using both improves the Exact Match metric, suggesting that the model generates more consistent answer spans.

7. Conclusion

We propose a new approach for question generation, relying on a BART Transformer based generative model, for which input data are enriched by semantic constraints. Question generation and answering are evaluated on several French corpora, and the whole approach is validated on a new corpus of digitized archive collection of a French Social Science journal. In particular we presented the question generation and question answering models that have been developed to adapt a Machine Reading model trained on Wikipedia to the *self-management* corpus of the ARCHIVAL project without any supervision.

8. Acknowledgements

This work has been partially funded by the Agence Nationale pour la Recherche (ANR) through the following program: ANR-19-CE38-0011 (ARCHIVAL).

This work was performed using HPC/AI resources from GENCI-IDRIS (Grant 2021-AD011012688).

The project leading to this publication has received funding from Excellence Initiative of Aix-Marseille - A*MIDEX (Archimedes Institute AMX-19-IET-009), a French "Investissements d'Avenir" Programme.

9. Bibliographical References

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics.

Béchet, F., Damnati, G., Heinecke, J., Marzinotto, G., and Nasr, A. (2017). CALOR-Frame : un corpus de textes encyclopédiques annoté en cadres sémantiques. In *ACor4French – Les corpus annotés du français - Atelier TALN*, Orléans, France, June.

Béchet, F., Aloui, C., Charlet, D., Damnati, G., Heinecke, J., Nasr, A., and Herledan, F. (2019). Calor-quest: generating a training corpus for machine reading comprehension models from shallow semantic annotations. In *MRQA: Machine Reading for Question Answering-Workshop at EMNLP-IJCNLP 2019-2019 Conference on Empirical Methods in Natural Language Processing*.

Bertsch, A. and Bethard, S. (2021). Detection of puffery on the english wikipedia. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 329–333.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

d'Hoffschmidt, M., Belblidia, W., Heinrich, Q., Brendlé, T., and Vidal, M. (2020). FQuAD: French question answering dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1193–1208, Online, November. Association for Computational Linguistics.

Du, X., Shao, J., and Cardie, C. (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.

Eddine, M. K., Tixier, A. J.-P., and Vazirgiannis, M. (2020). Barthez: a skilled pretrained french sequence-to-sequence model. *arXiv preprint arXiv:2010.12321*.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July. Association for Computational Linguistics.

Lyu, C., Shang, L., Graham, Y., Foster, J., Jiang, X., and Liu, Q. (2021). Improving unsupervised question answering via summarization-informed question generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4134–4148.

Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D., and Sagot, B. (2020). CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online, July. Association for Computational Linguistics.

Marzinotto, G., Auguste, J., Bechet, F., Damnati, G., and Nasr, A. (2018). Semantic frame parsing for information extraction : the calor corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*. European Language Resource Association.

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels, October. Association for Computational Linguistics.

Puri, R., Spring, R., Shoeybi, M., Patwary, M., and Catanzaro, B. (2020). Training question answering models from synthetic data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5811–5826.

Pyatkin, V., Roit, P., Michael, J., Goldberg, Y., Tsarfaty, R., and Dagan, I. (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, November. Association for Computational Linguistics.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (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.

Shakeri, S., dos Santos, C., Zhu, H., Ng, P., Nan, F., Wang, Z., Nallapati, R., and Xiang, B. (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.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2019). Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in Neural Information Processing Systems*, 32.

Zhang*, T., Kishore*, V., Wu*, F., Weinberger, K. Q., and Artzi, Y. (2020). Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.