A new approach to generate teacher-like questions guided by text spans extraction

Thomas Gerald, Sofiane Ettayeb, Louis Tamames, Ha Quang Le, Patrick Paroubek, Anne Vilnat

A new approach to generate teacher-like questions guided by text spans extraction

Thomas Gerald[1], **Sofiane Ettayeb**[1], **Louis Tamames**[1], **Ha-Quang**[3] **Le, Patrick Paroubek**[2], **Anne Vilnat**[2]

[1] Université Paris Saclay, CNRS, SATT Paris Saclay, LISN
[2] Université Paris Saclay, CNRS, LISN
firsname.lastname@lisn.upsaclay.fr
[3] stellia.ai

Abstract

Generating teacher-like questions and answers remains an open issue while being useful for students, teachers and teaching aid application providers. Given a textual course material, we are interested in generating non-factual questions which require an elaborate answer (implying some sort of analysis or reasoning). Despite the availability of annotated corpora of questions and answers, two main obstacles prevent the development of such generator using deep learning. Firstly, the amount of qualitative data is not sufficient to train generative approaches. Secondly, for a stand-alone application, we do not have an explicit support to guide the generation towards complex questions. In this article, we propose and compare several new retargetable language algorithms for answer text span support extraction and complex question generation, on secondary education course material use-case in French. We study the contribution of deep neural syntactic parsing and transformer based semantic representation, relying on the question type (according to our specific question typology) and the support text span in the context. We highlight the important role of nominal noun phrases and dependency relations, as well as the gain brought by recent transformer language models.

Keywords: corpus, question-answering, question generation

1. Introduction

In education, few mature approaches for teaching assistance using deep-learning methods are deployed. However, recent advances in Natural Language Processing (NLP) allows us to envision applications for the extraction, processing or generation of information for pedagogical purposes. We aim to develop a teaching assistant for generating non-factual questions (leading to elaborated answers) guided by text courses materials, implying some analysis or reasoning going beyond the simple restitution of factual data. The use-case chosen for our experiences concerns secondary courses of history in French language, in the context of a project funded by a technology transfer accelerator[1] in collaboration with a company[2] specialized in applications for education. The project aims to produce a question answering system with high pedagogical value inside an application able to guide students by partially answering course questions, redirecting them to relevant articles/courses, or proposing Multiple Choice Question (MCQ) to consolidate their knowledge.

To fulfill this objective we collected French question-answer pairs on education materials from both school books and Wikipedia. We currently have about 500 manually annotated question-answer pairs. Given a text course material, annotators qualified in the field of the course, were instructed to produce a set of questions of various types, indicating for each the text span of the course material from which the answer can be inferred.

The amount of annotated data we have now does not allow us to consider training or fine-tuning deep-learning generative approaches. Additionally, for a stand-alone application, we do not have access to the answer spans to guide the generation of the question, i.e. the passage of the text where the question generation system must focus on. In this paper we propose to train different transformer generation model for question generation using the corpus as evaluation material. Specifically, we study various support spans (named support or question support) extraction algorithms and compare them on their ability to produce teacher-like questions.

In the following, we first discuss works related to the question generation task; secondly, we describe the French educational corpus collected; in a third section we introduce the question support extraction algorithms and discuss our choices; then we present the experimental settings and protocol; we subsequently reports results of the experiments and discuss the abilities of the different approaches to generate teacher-like questions; finally, we conclude with a discussion of proposed and future approaches.

2. Related Works

Summaries, questions and answers generation have been and remain central topics in the NLP community. These different tasks have benefited from machine learning and deep learning advances. The "transformer" neural architecture (Vaswani et al., 2017) has provided significant improvements for generative approaches. These architec-

[1] *SATT Paris-Saclay, convention de maturation AVE-TAL*
[2] *stellia.ai, https://stellia.ai/*

tures have been revised in many ways by addressing multi-tasks (Raffel et al., 2020; Radford et al., 2019) or by scaling and increasing the size of the models and datasets used (Brown et al., 2020). Primarily developed for the English language these pre-trained models are now available in French with CamemBERT and FlauBERT (Martin et al., 2020; Le et al., 2020) language models (LM) or the BARThez generation model (Eddine et al., 2021). Most of the effective approaches now consider the multi-lingual settings for pre-training LM (Liu et al., 2020).

To adapt these models to a specific task, a common approach consists in fine-tuning language models on task oriented corpora. The corpus SQuAD (Rajpurkar et al., 2016) strongly participates in improving question-answering task, providing a large dataset of questions and extractive answers. More recently, Google published the corpus Natural Question (Kwiatkowski et al., 2019): a corpus with natural language questions, with long and short paragraphs for answers (extracted from the English Wikipedia). In conversational QA the corpus CANARD and QUAC (Elgohary et al., 2019; Choi et al., 2018) are available. For retrieval-based question-answering where documents are answers, the MSMarco passage dataset (Nguyen et al., 2016) is today the reference for training or fine-tuning models. If most QA corpora are available in English, French community also produced corpora such as FQuAD (Martin et al., 2020), Piaf (Keraron et al., 2020) or CALOR-QUEST (Bechet et al., 2019) for extractive QA. More recently, the CALOR-DIAL (Béchet et al., 2022) corpus addresses dialogue question answering for the French language. However, these corpora mainly rely on factual QA, where the answer is a short text such as a named entity, an event, a date, a quantity, or a location. Recently, a new corpus Autogestion (Antoine et al., 2022) has been created to address non-factual questions, the associated study demonstrates the inability of standard models to address most complex questions. All those corpora can be used for question generation (QG), answer generation, or answer extraction tasks.

Many QA works rely on those datasets particularly in machine reading comprehension (Liu et al., 2018; Yamada et al., 2020; Zhang et al., 2021). Moreover *QA* can be addressed within different frameworks such as the retrieval (Khattab and Zaharia, 2020; Karpukhin et al., 2020) or conversational (Anantha et al., 2021) one. Recent works have focused on explainable answers by Chain of Thought prompting (Wei et al., 2022) leveraging huge LM, similarly (Huang et al., 2022) proposed improvement of the approaches with no additional data needed. For QG, different kinds of approaches have been explored, such as the template-based approach where a pre-set of templates is filled with document information (Wolfe, 1976), the sequence-to-sequence approaches (Zi et al., 2019) or considering both (Fabbri et al., 2020). In a sequence to sequence model, additional information is usually given to guide the generation, the "question support". Extracting salient text spans is thus a key sub-task for text generation, it can be leveraged without any task prior, relying on part-of-speech extraction (Toutanvoa and Manning, 2000), dependency parsing (Surdeanu and Manning, 2010) or keyword extraction with KeyBERT (Grootendorst, 2020) using Bert embedding.

Although generation of either question or answer is getting closer from human writing, the lack of metrics still remains an issue. Generally in language generation the n-gram based approach such as BLEU (Papineni et al., 2002) or ROUGE (Lin, 2004) metrics are commonly involved. Some approaches using language model have been proposed, allowing to take into account meaning similarity instead of words similarity, such as the BERTScore (Zhang et al., 2020) based on embedding. Even if specific question generation metrics exist, as the Q-metric (Nema and Khapra, 2018), at the best of our knowledge none are as reliable as human judgment.

3. A French corpus for education

To gather a qualitative French corpus for education, we collect textbooks from middle and high school about History, Geography, Life science, and Civic Education from the "Livrescolaire"[3]. In addition, we retrieve Wikipedia articles related to this corpus. We filter them using Wikipedia API with queries based on the titles from the educational textbooks, then we bring together the subsections selected. We present to annotators paragraphs of a document and we asked them to create the following annotations:

• **A question:** written by the annotator.

• **The question type:** factual, descriptive, course or synthesis.

• **The question support(s):** extracted spans targeting the subject of the question.

• **Answer element(s):** the different passages allowing to answer the question.

• **The hand written answer:** from the annotator, using the answer elements.

Notice that for each document we asked annotators to create many annotations. We will focus here on the first three annotations. Translated examples are provided in table 1. We already launched two first annotation campaigns where

Type	Question	Support
Factual	In which year did Christopher Columbus reach America ?	Christopher Columbus reached America (1492)
Descriptive	What is a rotary press ?	A rotary press is a typographic press mounted on a cylinder, allowing continuous printing.
Course	How did the Europeans legitimize their domination?	Europeans rethink the hierarchy of people within a Christian and European-centered scheme which then serves to legitimize their domination
Synthesis	Why did some French people support the state of emergency after the 2015 Paris attacks ?	• protects them against the terrorist threat and the risk of a new attack, which is feared by all. • This exceptional regime continues to appear as "a necessity".

Table 1: Examples for the four question types

we obtained 412 questions. In the following, QAE will

[3]https://www.lelivrescolaire.fr/

refer to questions answers for education, with QAE-A the dataset obtained with teachers and QAE-B the dataset from a private annotation organism. In the future, the goal is to collect around 10.000 question/answer pairs in a final annotation campaign, which will provide a corpus to train deep neural network on complex question/answer generation.

4. Extracting the support to generate a question

Although a question support is available for generating questions and answers in the collected corpus, in real case this information is unavailable. In the following we define the automatic extraction of the question support.

4.1. Generate the question

In our experiments, our generative models take as input a context and a question support in order to generate a question. The context chosen is the paragraph from which a support is extracted. We make use of a special token $< hl >$ which is set around the question support. The format given is the following:

$$[pre\_context] < hl > [support] < hl > [post\_context]$$

For training, our cost function relies on minimizing the cross-entropy loss.

4.2. Extracting the question generation support

When building an automatic tool, the support for generating question is rarely given. In this study, we focus on the selection of such support by extracting specific information for each sentence of the original support extracted. For instance, given the sentence:
"After the failure of the Hungarian revolt during the Springtime of Nations, the Empire of Austria and its dynasty, the Habsburgs, reclaimed their full power"
***Original sentence**: Après l'échec de la révolution populaire hongroise lors du Printemps des peuples, l'Empire d'Autriche et sa dynastie, les Habsbourg, sont restaurés dans toute leur puissance.*
which part should be given to the model to generate the following question:
"What are the consequences of the Hungarian revolt failure during the Springtime of Nations?"
***Original sentence**: Quelle est la conséquence de l'échec de la révolution populaire hongroise?*
In the corpora available in the literature, a named entity, usually the answer, is given as the additional input in the form of a text span to guide the question generation system. However, a named entity is rarely sufficient to produce complex questions as it leads generation to only focus on a factual element. We thus try to automatically extract relevant text spans which would better guide the generation, using entities, syntactical units (such as objects of the predicate) or group of words standing together as a semantic unit (keyphrases). We studied approaches based on the following elements:

- **Source** (**SRC**): The question support selected by human annotators in our corpus, for instance: "the Empire of Austria and its dynasty, the Habsburgs, reclaimed their full power". In few cases, the selected support is not contiguous (annotators selected support in different paragraph), in this case one question by contiguous support will be generated leading to have many questions by annotation.
- **Named entities (ENT)** : A selection of named entities from any sentence overlapping with the source, for instance: "Springtime of Nations", "Empire of Austria", "Hasburgs".
- **Noun phrases (NP)** : A selection of noun phrases from any sentence overlapping with the source, for instance: "their full power", "its dynasty", "the Hungarian revolt". We did not take into account noun phrases overlapping with entities.
- **Object (OBJ)** : the object, i.e. the subtree annotated as OBJ by a dependency parser, from sentences overlapping with the source. For instance: "the Empire of Austria and its dynasty, the Habsburgs, reclaimed their full power" (here, it is the same as the source).
- **Keyphrase (KP)** : A selection of extracted "key passages" with a KeyBERT model (based on CamemBERT) from any sentence overlapping with the source. The KeyBERT model averages the embedding of the sentence and computes the cosine similarity of the contextual embeddings of text portions with this average. For each sentence we sample the top two key-phrases from 2 to 15 tokens using a diversity parameters of 0.6 (using Maximal Marginal Relevance). For contextual embeddings we used the *camenBERT-base* model[4] For the current example we obtained the following supports: "After the failure of the Hungarian revolt during the Springtime of Nations," and "reclaimed their full power"

We use the spacy [5] library with the *fr_core_news_lg* model to extract the support from the sentences. Notice that we use SRC support as default value.

5. Experimental settings

Model. In our experiments we fine-tuned three different models: **BARThez**[4] a French model designed for generative tasks having both encoder and decoder pretrained; **MBARThez**[4] model trained with objective similar to BARThez model using a multilingual setting (using the MBART architecture); **MBART**[4] model, a multilingual model trained on translation tasks in many languages. For multi-lingual approaches, we use a special token to specify the language of the input or output text. The code can be find on github[6]

Corpus. To train and validate our model we use the French datasets *Piaf* and *FQuAD*, for multilingual model the *SQuaD* datasets is additionally considered. We split the datasets to obtained the set described in table 2, and only use French datasets for validation. On the evaluation side, we use both our own educational corpus, QAE-A and

[4]https://huggingface.co
[5]https://spacy.io/
[6]https://github.com/tgeral68/EFRQA

Corpus	Train	Validation	Evaluation
SQuAD	87599	–	–
FQuAD	20731	1641	1547
Piaf	7375	1082	767
QAE-A	–	–	252
QAE-B	–	–	182

Table 2: Number of question and answer pairs in the different corpora. SQuAD has only been used for training.

Dataset	MBART	BARThez	MBARThez
FQuAD	41.8 / 90.9	42.4 / 91.0	**45.2 / 91.5**
Piaf	38.5 / 90.5	38.3 / 90.3	**39.7 / 90.6**
QAE-A	27.5 / 89.2	28.0 / **89.3**	**28.6 / 89.3**
QAE-B	35.4 / 90.4	37.3 / 90.5	**38.4 / 90.7**

Table 3: Results for the different models and dataset. We report the RougeL / BERTScore metrics.

QAE-B, and the test set of *Piaf* and *FQuAD*. For multilingual approaches, we duplicate each training corpus having the question translated (for FQuAD and Piaf the question is translated into English, for SQuAD into French) using the pre-trained MBART model.

Training. All models are fine-tuned using the same hyper-parameters: during the first 1000 iterations we linearly increase the learning rate to reach 1^{-4} (starting at 1^{-7}). The batch size is fixed to 128 samples using gradient accumulation. The context is truncated if exceeding 512 tokens. The optimizer is AdamW (Kingma and Ba, 2015); an epoch is specified to use 2000 batches and the training samples are randomly sampled. We stop the training if the RougeL value does not increase during 5 epochs (on validation set), the model with the best validation is saved and used in later experiments.

Evaluation. For the evaluation, we choose two metrics: RougeL, and BERTScore. **RougeL** [4], originally designed for machine translation, measures the number of n-grams shared between the prediction and the ground truth. We do not consider approaches based on geometrical mean like BLEU since the number of questions for each context may vary depending on the extraction approach. **BERTScore**[4] evaluates the similarity between two spans based on contextual embeddings extracted from a RoBERTa model[4] (xlm-roberta-large).

6. Results and analysis

In table 3 we report the performances of the different models for the two metrics. We observe that the MBARThez model outperforms the other models, including its monolingual counterpart. This result confirms the trend that increasing the number of parameters and exploiting the knowledge of datasets in different languages improve performance. The two models take advantage of being trained on generative tasks. Results for MBART demonstrate that translation models are not the best suited for question generation, the model being potentially biased by this initial task.

Although we did not report the side experiments where both the MBART and MBARThez models were fine-tuned on French corpora only, the validation results were lower, as could be expected. This emphasizes the improvement brought by training on datasets from different languages and demonstrate that we can augment the training sets with foreign corpora. We also note that BERTScore is less informative as values vary only within a small range and, in most cases, it follows the tendencies of the RougeL score. In the following, we will only consider RougeL and the MBARThez model.

6.1. Performances related to question support

In the following experiment, we evaluate performances based on the extracted question support. Table 4 reports the results obtained when considering the different question support. In addition to mean RougeL score, we process the mean for the best and worst rated questions, as for each source, several supports can be extracted and one question is generated per support, the mean number of supports extracted is also given. Extracting the object (**OBJ**) give the overall best (Mean in the table) generated question according to the four datasets while also maximizing the worst (Min) question generated. This extracting approach is thus reliable to generate valid questions and minimize the risk of generating poor or incorrect questions. The noun phrases (**NP**) maximize the best (Max in the table) generated questions. However, it is difficult to draw conclusions as a high number of supports were extracted for each source, and hence more questions were generated than for the other supports. Thus, extracting certain noun phrases allows us to get the questions most similar to those of the annotators. However, these results do not allow us to judge the quality of the questions, i.e. many other relevant questions can be produced for a same passage.

6.2. Human evaluation

If the previous metrics offer insights into how the models perform and can reproduce questions from the test set, we still lack a qualitative evaluation. We asked human annotators to evaluate the question quality. Each question was evaluated on a scale from 1 to 4 on three criteria: the syntax correctness, the meaning of the question, the answerability according to the context. The annotators also classified whether a question is factual or not.

Table 5 show the judgment of 5 annotators, with 30 questions each from QRE datasets. The best performances for syntax and question relevance are obtained with **OBJ** extraction, this reinforces the results of section 6.1.. However, for answerability, the **ENT** extraction achieves better results, it may be a consequence of the format of such questions (factual) and the format of the answer (unique entity). This intuition is supported by the factuality ratio (FACT), where only few (10%) questions are classified as factual. On the contrary, less factual questions are obtained through the **OBJ** extraction approach, which enhance its relevance to generate more complex questions.

Dataset	Sup	Mean	Max	Min	N
fquad	ENT	32.5	38.8	26.9	2.3
	NP	28.8	**45.7**	16.2	7.1
	KP	30.7	37.7	23.7	2.0
	OBJ	**33.6**	36.9	**30.6**	1.7
piaf	ENT	29.9	35.3	25.1	2.4
	NP	26.7	**40.9**	16.0	6.4
	KP	28.0	34.2	21.8	2.0
	OBJ	**31.5**	34.5	**28.7**	1.6
QAE-A	ENT	**25.9**	30.4	21.8	2.5
	NP	24.1	**38.1**	13.3	8.5
	KP	25.7	34.4	17.4	3.4
	OBJ	**25.9**	29.6	**22.8**	2.0
QAE-B	ENT	28.2	33.2	24.2	2.5
	NP	28.8	**43.6**	16.9	8.7
	KP	31.3	41.1	23.0	3.9
	OBJ	**32.9**	37.7	**28.9**	2.3

Table 4: MBARThez results for the different extraction (see section 4.2.). RougeL is reported for the average (**Mean**), maximum (**Max**) and minimum (**Min**) performance of each question grouped by source, with *N* indicating its mean number.

	COR	REL	ANS	FAC
SRC	3.63	3.17	3.17	60%/30%
ENT	3.53	2.73	**3.47**	83%/10%
NP	3.40	2.47	2.87	83%/ 7%
KP	3.60	2.60	2.83	83%/16%
OBJ	**3.67**	**2.87**	3.23	70%/**20**%
Total	3.57	2.77	3.11	75%/16%

Table 5: Human evaluation for the QRE-A & QRE-B datasets (on a scale from 1 to 4) with **COR** the syntax correctness, **REL** the question relevance, **ANS** the answerability and, **FAC** if the question is factual or not.

6.3. Performances according to question types

We show in table 6 the *RougeL* performances for the generated question from **SRC** support and **OBJ** support (as best performances are reached considering this last extraction approach). As a reminder, the course (COUR) and synthesis (SYNT) questions are usually more sophisticated, and thus harder to generate. The synthesis questions are particularly challenging since they often rely on different passages of the text. On the contrary, vocabulary questions are much easier to generate: firstly, because the associated text mostly relies on an explicit definition in the context, which limits the possibles questions, e.g. "Discrimination: treating someone differently because of their origin, skin color, religion, gender or sexual orientation, political or trade union orientation."; secondly, the questions are often simple and have few templates available, e.g. "What is discrimination?". Although we obtain the best performances within the manually annotated support (Mean column) on *RougeL* score, the results obtained considering the object of the sentence are not far behind, thus it remains a good candidate for generating this kind of questions. Interestingly, we observe better scores for **OBJ** based generation when looking at the average of maximum performances (Max column) for both courses and synthesis questions. We thus, empirically demonstrate that we can generate questions closer to the manually created ones and less factual using extraction of object sentences.

Dataset	QType	Mean	Max	Min	N
SRC (manually annotated support)					
QAE-A	FACT	26.2	”	”	1.0
	VOCA	**57.6**	”	”	1.0
	COUR	26.0	”	”	1.0
	SYNT	24.2	”	”	1.0
QAE-B	FACT	53.0	”	”	1.0
	VOCA	**53.7**	”	”	1.0
	COUR	35.0	35.5	34.4	1.1
	SYNT	19.3	23.9	15.3	2.4
OBJ (support based on sentence object extraction)					
QAE-A	FACT	22.1	23.5	20.6	1.5
	VOCA	**52.8**	**56.0**	**49.6**	1.4
	COUR	24.6	28.8	21.3	1.9
	SYNT	21.7	25.7	18.2	2.3
QAE-B	FACT	36.0	38.8	33.2	1.3
	VOCA	**48.1**	**51.9**	**44.6**	1.6
	COUR	33.6	36.8	30.3	1.6
	SYNT	18.3	27.0	12.3	4.2

Table 6: Results based on the different question type for support based on object and source (*RougeL* score).

7. Conclusion

We address the question generation task to generate teacher-like questions. To this end, we developed different algorithms and evaluated them on a new French corpus focused on educational question generation. Furthermore, we studied the impact of many support extraction methods in order to develop a reliable automatic question generation system for education. Then we spotlight the performances according the type and complexity of the question. We empirically demonstrate that it remains possible to improve the quality of questions using corpora in multiple languages (section 6.). Then, we compared different extraction methods to get the generation support. Even if human-extracted passages led to the best performances, we show that extracting sentence objects is a good candidate to automatically generate questions. Our human evaluation emphasized the relevance of object extraction, showing its ability to generate non factual question. Finally, we reported performances based on question types, showing that the difficulty of the questions fits our intuition, where reasoning questions are difficult to reproduce according to the current model training protocol and configuration. It is still difficult to automatically state the quality of the questions; we evaluate each generated question according to its likelihood with the annotated question which does not express

how good it is. In future work, we plan to explore evaluation metrics to better judge the quality of the generated questions. Furthermore, for complex questions, the answer may rely on several passages, the approaches designed here only take into account a single span for generating questions. This point will be addressed once the amount of collected data allows us to fine-tune transformers models (annotation of 10.000 questions-answers pairs is planned). Last but not least, according to the long-term objectives, the answer extraction/generation are foreseen, particularly to produce answers within an explanation scheme.

References

Anantha, Raviteja, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi, 2021. Open-domain question answering goes conversational via question rewriting. In *NAACL-HLT*. Association for Computational Linguistics.

Antoine, Elie, Jeremy Auguste, Frédéric Béchet, and Géraldine Damnati, 2022. Génération de questions à partir d'analyse sémantique pour l'adaptation non supervisée de modèles de compréhension de documents. *29e conférence sur le Traitement Automatique des Langues Naturelles (TALN)*.

Bechet, Frederic, Cindy Aloui, Delphine Charlet, Geraldine Damnati, Johannes Heinecke, Alexis Nasr, and Frederic Herledan, 2019. CALOR-QUEST : un corpus d'entraînement et d'évaluation pour la compréhension automatique de textes. In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019*.

Béchet, Frédéric, Ludivine Robert, Lina Rojas-Barahona, and Géraldine Damnati, 2022. Calor-Dial : a corpus for Conversational Question Answering on French encyclopedic documents. In *CIRCLE (Joint Conference of the Information Retrieval Communities in Europe)*. Samatan, France.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei, 2020. Language models are few-shot learners. *CoRR*.

Choi, Eunsol, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer, 2018. Quac: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Eddine, Moussa Kamal, Antoine J.-P. Tixier, and Michalis Vazirgiannis, 2021. Barthez: a skilled pretrained french sequence-to-sequence model. In *EMNLP (1)*.

Elgohary, Ahmed, Denis Peskov, and Jordan L. Boyd-Graber, 2019. Can you unpack that? learning to rewrite questions-in-context. In *EMNLP-IJCNLP*. Association for Computational Linguistics.

Fabbri, Alexander R., Patrick Ng, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang, 2020. Template-based question generation from retrieved sentences for improved unsupervised question answering. Association for Computational Linguistics.

Grootendorst, Maarten, 2020. Keybert: Minimal keyword extraction with bert.

Huang, Jiaxin, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han, 2022. Large language models can self-improve.

Karpukhin, Vladimir, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih, 2020. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*. Association for Computational Linguistics.

Keraron, Rachel, Guillaume Lancrenon, Mathilde Bras, Frédéric Allary, Gilles Moyse, Thomas Scialom, Edmundo-Pavel Soriano-Morales, and Jacopo Staiano, 2020. Project piaf: Building a native french question-answering dataset. In *Proceedings of The 12th Language Resources and Evaluation Conference*.

Khattab, Omar and Matei Zaharia, 2020. Colbert: Efficient and effective passage search via contextualized late interaction over BERT. In *SIGIR*. ACM.

Kingma, Diederik P. and Jimmy Ba, 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

Kwiatkowski, Tom, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov, 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*.

Le, Hang, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab, 2020. Flaubert: Unsupervised language model pre-training for french. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020*.

Lin, Chin-Yew, 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics.

Liu, Xiaodong, Yelong Shen, Kevin Duh, and Jianfeng Gao, 2018. Stochastic answer networks for machine reading comprehension. In *ACL (1)*. Association for Computational Linguistics.

Liu, Yinhan, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer, 2020. Multilingual denoising pre-training for neural machine translation.

Martin, d'Hoffschmidt, Vidal Maxime, Belblidia Wacim,

and Brendlé Tom, 2020. FQuAD: French Question Answering Dataset. *arXiv e-prints*.

Martin, Louis, Benjamin Müller, 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, ACL 2020, Online, July 5-10, 2020*.

Nema, Preksha and Mitesh M. Khapra, 2018. Towards a better metric for evaluating question generation systems. In *EMNLP*. Association for Computational Linguistics.

Nguyen, Tri, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng, 2016. MS MARCO: A human generated machine reading comprehension dataset. In *CoCo@NIPS*, volume 1773 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu, 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al., 2019. Language models are unsupervised multitask learners. *OpenAI blog*.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu, 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*.

Rajpurkar, Pranav, Jian Zhang, Konstantin Lopyrev, and Percy Liang, 2016. Squad: 100, 000+ questions for machine comprehension of text. In *EMNLP*. The Association for Computational Linguistics.

Surdeanu, Mihai and Christopher D. Manning, 2010. Ensemble models for dependency parsing: Cheap and good? In *HLT-NAACL*. The Association for Computational Linguistics.

Toutanvoa, Kristina and Christopher D. Manning, 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *EMNLP*. Association for Computational Linguistics.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems*.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou, 2022. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903.

Wolfe, John H., 1976. Automatic question generation from text - an aid to independent study.

Yamada, Ikuya, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto, 2020. LUKE: deep contextualized entity representations with entity-aware self-attention. In *EMNLP (1)*. Association for Computational Linguistics.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi, 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Zhang, Zhuosheng, Junjie Yang, and Hai Zhao, 2021. Retrospective reader for machine reading comprehension. In *AAAI*. AAAI Press.

Zi, Kangli, Xingwu Sun, Yanan Cao, Shi Wang, Xiaoming Feng, Zhaobo Ma, and Cungen Cao, 2019. Answer-focused and position-aware neural network for transfer learning in question generation. In *KSEM (2)*, Lecture Notes in Computer Science.