Sélectionner les ”bons” passages pour créer les ”bonnes” questions : Analyse et Évaluation d'un nouveau Corpus de Questions et Réponses pour l'Éducation

Thomas Gerald, Sofiane Ettayeb, Ha Quang Le, Gabriel Illouz, Patrick Paroubek, Anne Vilnat

Sélectionner les "bons" passages pour créer les "bonnes" questions : Analyse et Évaluation d'un nouveau Corpus de Questions et Réponses pour l'Éducation

Thomas Gerald*, Sofiane Ettayeb*, Ha Quang Le **, Gabriel Illouz***, Patrick Paroubek***, Anne Vilnat***

*Université Paris Saclay, CNRS, SATT Paris Saclay, LISN
prenom.nom@lisn.fr
**Professorbob.ai
ha-quang.le@polytechnique.edu
***Université Paris Saclay, CNRS, LISN
prenom.nom@lisn.fr

Résumé. Les systèmes intelligents pour le support scolaire sont aujourd'hui absents de la plupart des applications, alors que les récentes améliorations du Traitement Automatique des Langues (TAL) permettent d'imaginer des solutions innovantes. La création d'un système de questions-réponses reposant sur des sources scolaires permettrait d'accélérer, d'améliorer et de motiver l'apprentissage de l'étudiant. Dans ce contexte nous nous intéressons à la génération de questions au travers d'approches neuronales. Avec la récente création d'un corpus de questions-réponses par annotation de sources éducatives en langue française, nous disposons des ressources pour évaluer et développer de telles approches. Néanmoins, il faut considérer plusieurs obstacles : la quantité de données qualitatives n'est pas suffisante pour entraîner des approches génératives; dans le cadre d'une application autonome nous ne disposons pas explicitement du support pour la génération. Dans cette étude, nous proposons différentes méthodes d'extraction de ces supports comparant et analysant les résultats sur notre corpus et ceux de la littérature.

1 Introduction

Dans le milieu de l'enseignement, peu d'approches matures d'aide à l'apprentissage utilisant des algorithmes poussés d'apprentissage statistique sont aujourd'hui fonctionnelles. Les récentes avancées dans le traitement automatique des Langues (*TAL*) permettent pourtant de créer des outils pour traiter, extraire de l'information ou générer du contenu pour des documents et des supports de cours. Dans cet article, nous nous intéressons à cette problématique, en particulier en considérant les tâches de génération de questions et de création de réponses extractives ou abstractives. Nous travaillons en langue française afin de développer en partenariat avec l'industrie (entreprise ProfessorBob) un système automatique d'aide à l'enseignement. L'objectif poursuivi serait de fournir une application complète capable d'aider l'étudiant dans

son apprentissage, en répondant partiellement à ses questions de cours, en l'orientant sur les sujets à réviser ou en proposant des QCM. Aujourd'hui, nous proposons de mettre l'accent sur le système de questions-réponses, pour cela nous disposons de données en français que nous avons fait annoter. Ces données consistent en un ensemble de questions-réponses, avec pour chaque exemple une question rédigée par l'annotateur et une réponse correspondant à une partie sélectionnée dans le texte. Les documents sources sont extraits soit de manuels scolaires, soit de Wikipedia pour des sujets d'histoire, de géographie et d'éducation civique. Nous disposons d'environ 400 couples de question et réponse que nous compléterons dans de prochaines campagnes d'annotations. Il n'est donc aujourd'hui pas encore envisageable d'apprendre des modèles profonds génératifs sur ces données.

Dans l'application existante, afin de vérifier la pertinence des résultats, les questions et les réponses sont pré-générées, celles-ci pourront être filtrées afin d'être certains de ne pas produire d'erreurs. Pour évaluer la qualité des questions générées par le système, le volume des données nous permet seulement d'envisager

des approches *few-shot* ou *zero-shot*, sans entraînement sur les domaines cibles. Aussi, dans le cadre d'une application fonctionnelle nous ne disposons pas explicitement des parties de phrases ou de paragraphes cible à donner en entrée des modèles génératifs (support de la question). Pour cela, nous proposons d'étudier différents supports pour la génération de la question en se basant sur l'extraction de suite de mots pertinents dans les phrases cibles (via l'utilisation d'un arbre de dépendance). Enfin nous proposons d'étudier la pertinence des corpus obtenus, en mesurant les performances à partir des supports de génération relativement à une typologie des questions que nous avons établie.

2 Approches connexes

La génération dans le TAL La génération de résumés, de questions et de réponses sont des thèmes centraux dans le TAL. Ces différentes tâches ont profité des récentes avancées en apprentissage statistique, tout particulièrement grâce aux améliorations des approches neuronales (apprentissage profond). Avec les modèles “*transformeurs*” (Vaswani et al., 2017) et ses différentes configurations et améliorations (Radford et al., 2019; Raffel et al., 2020; Brown et al., 2020), la génération de texte a connu un regain d'intérêt. Ces architectures ont été adaptées pour la langue française dans plusieurs travaux (Eddine et al., 2021; Martin et al., 2020; Le et al., 2020) ou en multilingue (Winata et al., 2021). Nous allons utiliser ce type d'architecture afin de développer un modèle de génération de questions.

Génération de questions Nous nous intéressons plus particulièrement à la génération automatique de questions (**AQG**)

en utilisant les architectures susmentionnées.

Les données utilisées en **AQG** sont variées : du texte (Heilman et Smith, 2010), des images (Mostafazadeh et al., 2016) ou tout autre type de données structurées. Pour les données en langage naturel plusieurs approches existent, comme celles s'appuyant sur des modèles de questions (Wolfe, 1976), sur des modèles séquence vers séquence(Zi et al., 2019) ou les deux à la fois (Fabbri et al., 2020).

Toujours sur ces données textuelles, une des tâches est la génération de réponses avec comme entrée des questions.

Cette tâche est souvent assimilée à la tâche duale de celle de réponse aux questions (Chan et Fan, 2019). Les corpus de questions-réponses sont alors ré-utilisés en donnant en entrée le paragraphe utilisé pour la génération (que nous appellerons *contexte* par la suite) et la réponse à la question que l'on souhaite générer. Cela entraîne majoritairement la création de questions dites "factuelles" qui présentent un intérêt pédagogique limité (Dong et al., 2018). Des modèles n'utilisant pas la réponse en entrée ont également été développés (Lopez et al., 2021). Cependant, ces derniers ont tendance à générer des questions qui ont peu de rapport avec le paragraphe utilisé ou auxquelles les éléments de ce paragraphe ne suffisent pas pour apporter une réponse (Scialom et al., 2019).

Les entrées des modèles pour la génération Plusieurs travaux visent à fournir une entrée pertinente pour orienter la génération. Ainsi, il a été proposé de fournir un résumé automatique du contexte (Dugan et al., 2022), d'ajouter des meta-données comme des titres de paragraphes (Nguyen et al., 2022) ou de fournir les phrases contenant la réponse (Back et al., 2021). La langue des données est également un facteur important, les modèles multilingues pouvant profiter d'un *fine-tuning* sur des données plus nombreuses dans une langue secondaire (l'anglais) afin d'améliorer les performances sur la langue qui nous intéresse, ici le français (Kumar et al., 2019). Enfin la pertinence des questions par rapport à l'enseignement est un point clef. Ainsi, considérer seulement des questions amenant une réponse factuelle n'est pas souhaitable (Yao et al., 2012).

Corpus de questions réponses Les corpus de questions et réponses sont aujourd'hui nombreux et variés en terme de domaines. On discerne plusieurs configurations pour ces corpus telles que les approches en domaine ouvert (Fan et al., 2019; Kwiatkowski et al., 2019), incluant souvent des modèles de RI pour rechercher l'information, les approches en domaine restreint (Rajpurkar et al., 2016, 2018), où un contexte est donné dans lequel extraire la réponse), ou bien dans les systèmes de dialogue (Choi et al., 2018; Elgohary et al., 2019). Les deux premiers types de corpus, domaine ouvert et domaine restreint, sont exploitables pour la génération de questions, contrairement aux systèmes de dialogue où la question ne contient que rarement l'intégralité du contexte.

En langue française, il existe peu de corpus de questions et réponses disposant d'un nombre de données annotées suffisantes pour espérer pouvoir entraîner ou adapter des modèles neuronaux. Les plus connus sont FQuAD (Martin et al., 2020) et Piaf (Keraron et al., 2020) contenant respectivement 23919 et 9224 couples de questions et réponses. Bien que disposant d'un grand nombre d'exemples d'entraînement, les réponses sont majoritairement factuelles. Le corpus CALOR-QUEST fournit en français une base de questions et réponses générées de manière semi-supervisée (Bechet et al., 2019). Plus récemment, CALOR-DIAL intègre les questions et réponses dans les systèmes de dialogue (Béchet et al., 2022).

3 Un corpus pour la génération et la sélection de questions réponses

Les sources. Pour parvenir à un corpus regroupant ces qualités, nous avons récolté plusieurs supports de cours dans différentes matières (Histoire, Éducation Civique et Morale, Géogra-

phie et Science et Vie de La Terre). Aussi, plusieurs niveaux scolaires sont considérés allant de la classe de 6^e à la classe de 1^{re}. Les données sont extraites des livres numériques de la plateforme “Lelivrescolaire” [1]. En supplément de cette ressource, nous utilisons plusieurs articles Wikipedia. Ces derniers sont filtrés avec le moteur de recherche fourni par wikipedia en utilisant comme requêtes les entités nommées extraites des supports scolaires ; les sous-sections sont proposées comme documents à annoter sur la plateforme.

Les annotations. Lors de l'annotation nous récupérons les éléments suivants :

1. **La question** : une question rédigée par l'annotateur portant sur le document ;
2. **Le type de la question** : nous proposons quatre classes de question différentes : factuelles, descriptives, de cours et d'agrégation de l'information ;
3. **Le support de la question** : un passage dans le texte servant de support à la génération d'une question ;
4. **Les éléments de réponse** : une sélection de plusieurs passages permettant de répondre aux différents éléments de la question ;
5. **La réponse rédigée** : une réponse écrite par l'annotateur reprenant les différents éléments de la réponse extraite ;

Des exemples de ces annotations sont fournies Table 1. Dans cette étude nous nous concentrons sur les trois premiers éléments.

Les campagnes. Lors de deux campagnes d'annotations expérimentales menées en 2022 nous avons obtenu 412 questions réponses réparties sur plusieurs documents. Aujourd'hui les couples questions-réponses ne disposent pas toujours de réponses rédigées. Dans la suite, nous nommerons ainsi les corpus associés aux campagnes : Question Réponse pour L'enseignement (*QRE*), avec *QRE-A* le dataset obtenu avec des professionnels de l'éducation et *QRE-B* les annotations obtenues avec le second groupe via une structure d'annotations. Nous planifions la récolte d'environ 10000 couples de questions et de réponses dans la prochaine campagne d'annotation afin de collecter une quantité de données suffisante pour l'apprentissage de réseaux de neurones profonds.

Type	**Question**	**Support**
Factuelle	En quelle année Christophe Colomb atteint l'Amérique ?	Christophe Colomb atteint l'Amérique (1492)
Descriptive	Qu'est-ce qu'une rotative ?	Rotative : presse typographique montée sur un cylindre, permettant d'imprimer en continu.
Cours	Comment les Européens ont légitimé leur domination ?	Les Européens repensent la hiérarchie des peuples au sein d'un schéma chrétien et européocentré qui sert ensuite à légitimer leur domination
Synthèse	Pourquoi certains français ont-ils soutenu l'Etat d'urgence après les attentats de Paris de 2015 ?	• les protège contre la menace terroriste et le risque d'un nouvel attentat, redouté de tous. • ce régime d'exception continue d'apparaître comme « une nécessité »

TAB. 1 – *Exemples d'annotations obtenues pour les différents types de questions*

1. https://www.lelivrescolaire.fr/

4 Génération de questions et graine de génération

Dans les corpus annotés, nous disposons d'un support extrait du texte qui nous aide à guider la génération (il s'agit généralement de la réponse à la question). Dans les cas applicatifs, cette donnée n'est pas disponible et il faut alors déterminer quels sont les meilleurs passages du texte pour générer la question. Afin de générer une question, nous utilisons un modèle "transformeur" de type Seq2Seq.

4.1 La génération de question

Pour générer une question nous donnons en entrée du modèle génératif une chaîne de caractères, le contexte, qui est représenté par un paragraphe. Afin de diriger la question, une information supplémentaire est fournie, le support, qui est une partie du contexte. Le support est entouré par un caractère spécial $< hl >$. Le format de donnée en entrée est le suivant :

$$[contexte\_antrieur] < hl > [support] < hl > [context\_postrieur]$$

Pour l'apprentissage du modèle nous utilisons la fonction de coût proposé dans l'implémentation du modèle s'appuyant sur la minimisation de l'entropie croisée.

4.2 Les différentes approches d'extractions de graine de génération

Dans le cas d'une application autonome, le support n'est pas explicitement fourni. Dans cette étude, nous nous intéressons tout particulièrement à la sélection de celui-ci, l'objectif étant de produire la question proposée dans les corpus. Par exemple, étant donné le contexte :
"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."
le but est de déterminer la partie à extraire pour générer la question suivante :
"Quelle est la conséquence de l'échec de la révolution populaire hongroise ?"
Dans les corpus de la littérature, c'est souvent une entité nommée qui est ciblée dans le texte, qui, de plus, correspond à la réponse.

Dans les corpus que nous avons collectés, le support de la question est rarement une entité, du fait du domaine pédagogique.

Dans un système automatique, l'extraction du support de la question est nécessaire, il s'agit donc de retrouver les potentielles réponses.On se propose d'étudier plusieurs approches pour l'extraction du support :

- **Source** : Le texte sélectionné comme support de la question pour notre corpus ou comme réponse pour les corpus *FQuAD* et Piaf, par exemple : "l'Empire d'Autriche et sa dynastie, les Habsbourg, sont restaurés dans toute leur puissance",
- **les entités nommées (ENT)** : Une sélection des entités des phrases où se trouve la réponse. Par exemple : "Printemps des peuples", "l'Empire d'Autriche", "Habsbourg" ont été automatiquement extraites.
- **les groupes nominaux (GN)** : Une sélection des groupes nominaux des phrases où se trouve la réponse, les groupes nominaux sélectionnés ne devant pas être présents dans les entités retrouvés. Par exemple les groupe nominaux "l'échec", "la révolution populaire", "sa dynastie" et "toute leur puissance" ont été automatiquement extraits.

- **Les compléments d'objet (CO)** : L'objectif est ici de retrouver le complément d'objet, pour cela nous extrayons les rôles de dépendances OBJ, IOBJ et extrayons le sous arbre de dépendance complet associé, ici nous obtenons "l'Empire d'Autriche et sa dynastie, les Habsbourg, sont restaurés dans toute leur puissance"
- **Phrases (PH)** : Dans ce dernier cas, nous sélectionnons l'intégralité des phrases en intersection avec la réponse ; ici nous obtiendrons : “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.”

Pour extraire cette information, nous considérons les phrases de la réponse pour *Piaf* et *FQuAD* ou de la source sélectionnée par les annotateurs pour les deux corpus *QRE*, nous laissons les problématiques de recherche d'information (phrase) pour des travaux ultérieurs. Dans certains cas, aucun support supplémentaire n'est trouvé. Dans ce cas-là nous utilisons par défaut la source de la question. Dans tous les cas, nous extrayons les passages sur les phrases sélectionnées en support de question (resp en réponse pour FQuAD et Piaf).

5 Protocole expérimental

Les données. Pour l'entraînement et la validation de nos approches nous utilisons les corpus *Piaf* (Keraron et al., 2020) et *FQuAD* (Martin et al., 2020). Nous découpons le dataset Piaf en ensemble d'entraînement, de validation et d'évaluation. Pour *FQuAD*, un ensemble d'apprentissage et un ensemble de validation sont présents, nous découpons l'ensemble de validation afin d'obtenir un nouvel ensemble de validation et un ensemble d'évaluation. La taille de chacun des ensembles est décrite dans la table 2.

Pour évaluer les performances nous utilisons conjointement aux ensembles d'évaluation de *FQuAD* et *Piaf* les corpus *QRE-A* et *QRE-B* obtenus sur les supports scolaires lors des deux premières campagnes d'annotations.

Corpus	Entraînement			Validation			Évaluation		
	DOC	PAR	QUE	DOC	PAR	QUE	DOC	PAR	QUE
FQuAD	117	4921	20731	9	405	1641	9	363	1547
Piaf	428	1478	7375	92	217	1082	91	154	767
QRE-A	–	–	–	–	–	–			252
QRE-B	–	–	–	–	–	–			182
Total	545	6399	28106	101	622	2723	100	517	2726

TAB. 2 – *Les tailles des ensembles d'entraînement, de validation et d'évaluation pour les corpus FQuAD et Piaf et d'évaluation pour les corpus récoltés QRE-A et QRE-B.* DOC *est le nombre de documents,* PAR *le nombre de paragraphes,* QUE *le nombre de questions*

Évaluation. Pour évaluer les expériences, nous avons sélectionné deux métriques :

- **rougeL** (Lin, 2004) : Métrique originellement conçue pour l'évaluation de la traduction automatique mesurant le nombre de n-grams (nombre de tokens) partagés entre la source et la prédiction. Cette métrique est préférée aux approches similaires comme

BLEU où la performance globale est calculée via une moyenne géométrique, ici nous voulons être en mesure d'évaluer des sous-parties du corpus.

- **BERTScore** (Zhang et al., 2020) : La métrique BERTScore calcule la similarité entre deux paragraphes en regardant non pas les mots similaires mais les plongements contextuels obtenus via un modèle BERT.

Pour ces deux métriques nous reportons les scores multipliés par 100 pour favoriser la lisibilité dans les tables 3 et 5.

Entraînement et génération. Pour l'entraînement nous utilisons un modèle T5 pré-entraîné en français [2]. Nous sélectionnons la meilleure itération du modèle en accord avec les performances obtenues sur les ensembles de validation de *Piaf* et *FQuAD* (voir table 2). Le modèle est appris sur les réponses extraites du texte données pour chaque question, plusieurs réponses étant disponibles pour chaque annotation, nous tirons uniformément une réponse.

Pour l'optimisation nous avons sélectionné la méthode Adam (Kingma et Ba, 2015), en utilisant un *learning-rate* de $1e-4$ avec un *warmup* linéaire (sur 500 itérations démarrant à $5e-7$) et une taille de *batch* de 128 (avec accumulation du gradient). Pour la génération des questions nous utilisons *generate* proposée dans la bibliothèque *HuggingFace* en fixant le nombre maximal de tokens à 64. Nous expérimentons sur 4 types de graines de génération identifiés ci-dessus. Pour obtenir ces graines, nous utilisons la bibliothèque *spacy* [3] avec le modèle *fr_core_news_lg*.

6 Résultats et Analyses

Dans un premier temps, nous reportons les résultats obtenus dans la table 3 en considérant le support de la question sélectionné par les annotateurs. Notons que pour le dataset *QRE-B* plusieurs passages non contigus pouvant être sélectionnés, dans ce cas nous générons une question par passage. On remarquera tout d'abord que les performances obtenues sur nos deux collections sont inférieures en moyenne à celles obtenus sur FQuAD et PIAF. Nous pouvons supposer que ce résultat est dû aux différences entre les domaines (syntaxique, lexical) ou aux différences de pertinence de support entre les corpus. La différence entre *QRE-A* et *QRE-B* semble montrer que les questions sont plus difficiles pour *QRE-A* ou bien que la sélection ne correspond pas à ce que le modèle attend en entrée. Des expériences supplémentaires permettront d'infirmer ou d'affirmer cette dernière hypothèse.

6.1 Les différents supports

Dans la table 4, nous reportons les résultats pour les différents supports de question sélectionnés. Comme plusieurs supports peuvent être retrouvés pour une même question (plusieurs COD/COI/entités dans les phrases de la source), nous calculons la moyennes des scores par question, la moyenne des scores les plus élevés pour une question, la moyenne des scores les plus faibles pour une question ainsi que le nombre moyen de support par question, noté N.

Les résultats obtenus sur *QRE-B* sont proches de ceux que l'on peut obtenir sur les datasets de référence (FQuAD et Piaf), il est donc possible de trouver un support de génération pour

2. https ://huggingface.co/airKlizz/t5-base-multi-fr-wiki-news
3. https ://spacy.io/

Sélectionner les “bons” passages pour créer les “bonnes” questions

	Moyenne		N
	rougeL	BERTScore	N
FQuAD	42.0	90.8	1
PIAF	36.4	90.0	1
QRE-A	25.0	90.0	1
QRE-B	34.2	90.1	1.4

TAB. 3 – *Résultats pour la génération de questions en utilisant la source. N est le nombre de support de question moyen pour une question.*

Dataset	Sup	Moyenne		Maximum		Minimum		N
		rougeL	BERTS	rougeL	BERTS	rougeL	BERTS	
FQuAD	ENT	31.4	89.0	37.6	90.1	26.0	88.0	2.3
	GN	28.1	88.5	**44.1**	**91.2**	15.9	86.0	7.1
	CO	**32.4**	**89.2**	35.8	89.8	**29.3**	**88.6**	1.7
	PH	27.2	88.5	27.4	88.6	27.0	88.5	1.0
Piaf	ENT	28.9	88.6	34.4	89.6	24.0	87.6	2.4
	GN	25.7	88.1	**39.6**	**90.5**	15.0	85.9	6.4
	CO	**30.3**	**88.7**	33.1	89.3	**27.7**	**88.2**	1.6
	PH	25.4	88.1	25.6	88.1	25.3	88.0	1.0
QRE-A	ENT	23.5	88.1	27.7	88.9	20.1	87.2	2.5
	GN	22.1	88.0	**35.1**	**90.3**	12.5	85.6	8.5
	CO	23.2	88.1	26.5	88.7	20.1	87.4	2.0
	PH	**25.4**	**88.5**	29.9	89.3	**21.2**	**87.6**	2.0
QRE-B	ENT	25.2	88.4	30.2	89.3	20.9	87.4	2.5
	GN	25.7	88.6	**40.4**	**91.1**	14.9	86.2	8.7
	CO	**29.9**	**89.1**	34.0	89.9	**26.2**	**88.3**	2.3
	PH	26.8	88.9	29.6	89.4	24.4	88.4	2.0

TAB. 4 – *Résultats pour rougeL et BERTScore sur les corpus d'évaluation et les différents supports de question. N est le nombre de support de question moyen pour une question.*

notre corpus qui soit adapté. Notons que le corpus *QRE-B* n'a pas été annoté par des personnes formées pour l'éducation contrairement au corpus *QRE-A*. Les résultats obtenus sur *QRE-A* sont généralement moins bons, probablement du fait de la complexité des questions produites par les professionnels. On remarquera aussi que le meilleur support de question (en moyenne) repose majoritairement (non vérifié pour *QRE-A*) sur l'extraction des compléments d'objet, aussi il maximise dans la plupart des cas les plus faibles résultats ; il s'agit donc d'un support intéressant afin de minimiser le risque de génération de mauvaises réponses.

Pour le corpus *QRE-A* nous obtenons des résultats légèrement supérieurs à ceux obtenus dans le cas du support annoté, ce qui montre que nous pouvons proposer un support de génération de manière automatique sans effet négatif sur les performances (ce qui ne signifie pas pour autant de bonnes questions). Notons enfin que contrairement aux autres corpus, le support par extraction de phrase permet d'obtenir les plus hauts résultats. Plusieurs causes peuvent être à l'origine de cet effet : des questions moins précises ou des types de questions difficiles. Notons

Dataset	QType	Moyenne		Maximum		Minimum		N
		rougeL	BERTS	rougeL	BERTS	rougeL	BERTS	
QRE-A	TOTAL	22.79	88.07	39.96	90.99	10.51	84.93	15.9
	FACT	23.74	88.51	43.84	91.68	11.35	**85.25**	14.34
	VOCA	**35.43**	**89.44**	**62.62**	**93.96**	**13.19**	85.13	13.11
	COUR	22.63	88.15	40.15	91.21	10.38	85.16	16.21
	SYNT	19.73	87.58	33.57	89.98	9.74	84.67	16.82
QRE-B	TOTAL	27.05	88.77	47.44	92.25	11.61	85.35	16.9
	FACT	**32.92**	**89.77**	**56.64**	**93.64**	**13.56**	85.79	10.1
	VOCA	31.39	88.84	55.11	92.6	12.72	85.67	11.68
	COUR	26.81	89.16	44.39	92.31	12.38	**85.92**	12.7
	SYNT	19.57	87.53	37.92	90.85	8.49	84.15	30.41
FQUAD	TOTAL	30.11	88.86	51.19	92.29	13.11	85.31	13.12
PIAF	TOTAL	27.79	88.4	46.21	91.64	12.53	85.18	12.37

TAB. 5 – *Résultats pour les différents types de questions, FACT (factuelles), VOCA (vocabulaire), COUR (cours) et SYNT (raisonnement et synthèse). Les résultats reportés cumulent les sélections (**Source**, **ENT**, **GN**, **CO**,**PH**).*

que des expériences et des évaluations humaines supplémentaires permettraient de mieux trancher. Enfin, en observant les résultats de la colonne “Maximum”, pour la totalité des corpus, les groupes nominaux permettent d’obtenir les meilleures performances. Ce résultat est dû au grand nombre de supports proposés offrant plus de variété dans les questions générées. Il existe donc dans les textes des groupes nominaux permettant de créer une question similaire à celle de l’annotateur.

6.2 Les types des questions

Dans le tableau 5 nous nous intéressons aux performances obtenues en lien avec le type de questions. Pour rappel, intuitivement les questions les plus difficiles sont celles des types cours et synthèse, les performances reportées appuient cette hypothèse. Sur les questions factuelles les résultats diffèrent selon les corpus. Sur *QRE-A* celles-ci semblent plus difficiles à générer. Nous avons observé dans ce corpus des questions doubles pouvant engendrer cette différence de performances observés. En revanche pour les deux corpus, les questions de vocabulaire sont plus facilement obtenues ; cela s’explique par la présence dans le corpus annoté, d’encadrés regroupant des listes de définitions, ces contenus donnent naturellement lieu à des questions de vocabulaire telles que "Quelle est la définition..."

7 Conclusion

Dans cet article nous avons étudié les problématiques de la génération de question sur deux corpus que nous avons récemment collectés. Nous avons mis en place un protocole afin d’étudier les différents supports de génération de la question. Nous montrons expérimentalement, la pertinence des différentes méthodes d’extraction de support et mesurons leur adéquation

à nos données. Nous montrons expérimentalement la pertinence de la typologie des questions produites et la cohérence des pré-campagnes réalisées. Néanmoins plusieurs points sont à améliorer ou à éclaircir : 1) Les métriques d'évaluations ne rendent pas compte de la pertinence et de la qualité des questions générées mais seulement de leur similarité avec la question manuellement produite ; 2) Nous avons mené des expériences sur un unique modèle avec une configuration monolingue, des expériences préliminaires démontrent l'efficacité des modèles multilingues ; 3) Les performances pour les questions difficiles (cours et synthèse) sont généralement faibles, pour cela nous travaillons sur la réalisation d'un corpus similaire de taille conséquente (environ 10000 couples questions réponses) afin de pouvoir entraîner des modèles de bout-en-bout.

Avec cette étude, nous posons les fondations pour la création d'un système de questions-réponses pour l'éducation. Les corpus et matériaux utilisés seront mis à disposition de la communauté dans un futur proche.

Références

Back, S., A. Kedia, S. C. Chinthakindi, H. Lee, et J. Choo (2021). Learning to generate questions by learning to recover answer-containing sentences. In *Findings of the Association for Computational Linguistics : ACL-IJCNLP 2021*.

Bechet, F., C. Aloui, D. Charlet, G. Damnati, J. Heinecke, A. Nasr, et F. 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, F., L. Robert, L. Rojas-Barahona, et G. 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, T. B., B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, et D. Amodei (2020). Language models are few-shot learners. *CoRR*.

Chan, Y.-H. et Y.-C. Fan (2019). A recurrent BERT-based model for question generation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*.

Choi, E., H. He, M. Iyyer, M. Yatskar, W. Yih, Y. Choi, P. Liang, et L. Zettlemoyer (2018). Quac : Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Dong, X., Y. Hong, X. Chen, W. Li, M. Zhang, et Q. Zhu (2018). Neural question generation with semantics of question type. In *Natural Language Processing and Chinese Computing*.

Dugan, L., E. Miltsakaki, S. Upadhyay, E. Ginsberg, H. Gonzalez, D. Choi, C. Yuan, et C. Callison-Burch (2022). A feasibility study of answer-agnostic question generation for education. In *Findings of the Association for Computational Linguistics : ACL 2022*.

Eddine, M. K., A. J. Tixier, et M. Vazirgiannis (2021). Barthez : a skilled pretrained french sequence-to-sequence model. In *EMNLP (1)*.

Elgohary, A., D. Peskov, et J. L. Boyd-Graber (2019). Can you unpack that ? learning to rewrite questions-in-context. In *EMNLP-IJCNLP*. Association for Computational Linguistics.

Fabbri, A. R., P. Ng, Z. Wang, R. Nallapati, et B. Xiang (2020). Template-based question generation from retrieved sentences for improved unsupervised question answering. Association for Computational Linguistics.

Fan, A., Y. Jernite, E. Perez, D. Grangier, J. Weston, et M. Auli (2019). ELI5 : Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Heilman, M. et N. A. Smith (2010). Good question ! statistical ranking for question generation. In *The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Keraron, R., G. Lancrenon, M. Bras, F. Allary, G. Moyse, T. Scialom, E.-P. Soriano-Morales, et J. Staiano (2020). Project piaf : Building a native french question-answering dataset. In *Proceedings of The 12th Language Resources and Evaluation Conference*.

Kingma, D. P. et J. Ba (2015). Adam : A method for stochastic optimization. In *ICLR (Poster)*.

Kumar, V., N. Joshi, A. Mukherjee, G. Ramakrishnan, et P. Jyothi (2019). Cross-lingual training for automatic question generation.

Kwiatkowski, T., J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, K. Toutanova, L. Jones, M. Kelcey, M.-W. Chang, A. M. Dai, J. Uszkoreit, Q. Le, et S. Petrov (2019). Natural questions : A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*.

Le, H., L. Vial, J. Frej, V. Segonne, M. Coavoux, B. Lecouteux, A. Allauzen, B. Crabbé, L. Besacier, et D. Schwab (2020). Flaubert : Unsupervised language model pre-training for french. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020*.

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

Lopez, L. E., D. K. Cruz, J. C. B. Cruz, et C. Cheng (2021). Simplifying paragraph-level question generation via transformer language models. In *PRICAI (2)*. Springer.

Martin, d., V. Maxime, B. Wacim, et B. Tom (2020). FQuAD : French Question Answering Dataset. *arXiv e-prints*.

Martin, L., B. Müller, P. J. O. Suárez, Y. Dupont, L. Romary, É. de la Clergerie, D. Seddah, et B. 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*.

Mostafazadeh, N., I. Misra, J. Devlin, M. Mitchell, X. He, et L. Vanderwende (2016). Generating natural questions about an image. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.

Nguyen, H. A., S. Bhat, S. Moore, N. Bier, et J. Stamper (2022). Towards generalized methods for automatic question generation in educational domains. In *Educating for a New Future : Making Sense of Technology-Enhanced Learning Adoption*.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*.

Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, et P. J. Liu (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*.

Rajpurkar, P., R. Jia, et P. Liang (2018). Know what you don't know : Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

Rajpurkar, P., J. Zhang, K. Lopyrev, et P. Liang (2016). Squad : 100, 000+ questions for machine comprehension of text. In *EMNLP*. The Association for Computational Linguistics.

Scialom, T., B. Piwowarski, et J. Staiano (2019). Self-attention architectures for answer-agnostic neural question generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, et I. Polosukhin (2017). Attention is all you need. In *Advances in Neural Information Processing Systems 30 : Annual Conference on Neural Information Processing Systems*.

Winata, G. I., A. Madotto, Z. Lin, R. Liu, J. Yosinski, et P. Fung (2021). Language models are few-shot multilingual learners. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*.

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

Yao, X., G. Bouma, et Y. Zhang (2012). Semantics-based question generation and implementation. *Dialogue Discourse 3*, 11–42.

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

Zi, K., X. Sun, Y. Cao, S. Wang, X. Feng, Z. Ma, et C. Cao (2019). Answer-focused and position-aware neural network for transfer learning in question generation. In *KSEM (2)*, Lecture Notes in Computer Science.

Summary

Intelligent systems for teaching and learning assistance are missing from most applications, even though recent improvements in NLP allow us to imagine innovative solutions. The creation of a question-answer system based on academic sources would accelerate, improve and motivate student learning. In this context, we are interested in the generation of questions through neural approaches. With the recent annotation of a corpus of questions and answers in French, we have the resources to evaluate and develop such approaches. Nevertheless, several obstacles must be considered: the amount of qualitative data is not sufficient to train generative approaches; in the context of a stand-alone application we do not explicitly have the support for generation. In this study we propose different methods for extracting these supports comparing and analyzing the results on our corpus and those of the literature.