Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles
Avignon, France, 27 juin au 1er juillet 2022

Génération de questions à partir d'analyse sémantique pour l'adaptation non supervisée de modèles de compréhension de documents

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

RÉSUMÉ

La génération automatique de questions à partir de textes peut permettre d'obtenir des corpus d'apprentissage pour des modèles de compréhension de documents de type question/réponse sur des textes. Si cette tâche de génération est désormais appréhendée par des modèles de type séquence-à-séquence basés sur de grands modèles de langage pré-entraînés, le choix des segments réponses à partir desquels seront générées les questions est l'un des principaux aspects différenciant les méthodes de génération de corpus de question/réponse. Nous proposons dans cette étude d'exploiter l'analyse sémantique de textes pour sélectionner des réponses plausibles et enrichir le processus de génération par des traits sémantiques génériques. Les questions générées sont évaluées dans leur capacité à être utilisées pour entraîner un modèle de question-réponse sur un nouveau corpus d'archives numérisées.

ABSTRACT

Question generation from semantic analysis for unsupervised adaptation of document understanding models

Question generation can be used as a way to provide enhanced training material for Machine Reading Question Answering algorithms.If this generation task is handled by seq2seq models based on large pre-trained language models, the choice of answer segments from which questions are generated is one of the most differentiating aspect among the various approaches. We propose in this study to exploit semantic analysis of texts to select plausible answers and enrich the generation process by generic semantic features. The generated questions are evaluated in their capacity to be used for training a question answering model on a new corpus of digitized archives.

MOTS-CLÉS : Génération de questions, Compréhension de documents, Question/Réponse, Humanités numériques.

KEYWORDS: Question Generation, Machine reading Question Answering, Digital Humanities.

1 Introduction

La compréhension automatique de documents et la génération de questions à partir de textes sont deux tâches *miroirs* du Traitement Automatique de la Langue (TAL). Traditionnellement traitées par des *pipelines* complexes basés sur des modèles très différents, recherche d'information pour les modèles de question/réponse et analyse syntaxique pour la génération de questions, elles ont

Volume 1 : conférence principale, pages 104–115.

récemment été unifiées depuis l'avènement des méthodes *end-to-end* basées sur des modèles de langage pré-entraînés.

Ainsi, comme présenté dans (Du *et al.*, 2017), la génération de questions peut être modélisée comme une tâche de génération de texte où un modèle de type séquence-à-séquence est entraîné à *traduire* une séquence de mots représentant une phrase ou un passage en une autre séquence de mots représentant une question, sans passer par une analyse linguistique explicite comme c'était le cas auparavant. De son côté la compréhension automatique de documents peut être vue comme une tâche d'étiquetage consistant à apprendre, à partir d'un couple *(texte,question)*, quels mots devaient être étiquetés avec les labels *début de réponse* et *fin de réponse* dans le texte.

Le développement de modèles de langage pré-entraînés utilisés conjointement avec de grands corpus de triplets *question/réponse/contexte* tels que *SQuAD* (Rajpurkar *et al.*, 2016) peuvent être utilisés pour entraîner directement à la fois des modèles de traduction *réponse-vers-question* et des modèles d'étiquetage *question-vers-réponse*. Si les performances de ces modèles sont impressionnantes sur ces corpus de références, contenant généralement du texte issu de *Wikipedia* et des questions *simples* obtenues par *crowdsourcing*, la généralisation de ces modèles à des corpus contenant des textes plus complexes et à des questions moins littérales reste un défi. C'est dans ce cadre que nous proposons dans cette étude une méthode d'adaptation non supervisée de modèles de compréhension de texte basée sur la génération automatique de questions.

Les contributions de cette étude se situent à deux niveaux : d'une part la comparaison de différentes méthodes d'encodage d'informations sémantiques pour la génération de questions évaluées par rapport à leur capacité à entraîner un modèle de question/réponse sur un nouveau corpus de textes ; d'autre part l'étude de la capacité de généralisation de modèles de compréhension et de génération de question sur un nouveau corpus contenant des textes et des questions plus complexes que celles pouvant se trouver dans les corpus de références tels que *SQuAD* (Rajpurkar *et al.*, 2016) sur l'anglais ou *FQuAD* (d'Hoffschmidt *et al.*, 2020) pour le français.

2 Génération de questions

Plusieurs approches ont été proposées pour utiliser des triplets synthétiques *question/réponse/contexte* afin d'entraîner des modèles de lecture automatique dans une perspective d'augmentation des données ou dans des configurations d'adaptation non supervisée de type *zero-shot*. Si à chaque fois ce sont des modèles de génération de type traduction *réponse-vers-question* qui sont utilisés, les approches diffèrent par la méthode de sélection des *réponses* potentielles dans un texte à partir desquels les questions seront générées. Ainsi (Puri *et al.*, 2020) sélectionnent les réponses candidates avec un modèle de détection à base de *Transformers* et génère les questions correspondantes avec le modèle de génération *GPT-2*. Dans une approche plus systématique, (Shakeri *et al.*, 2020) prédisent des paires *(question, réponse)* à partir d'un passage en considérant systématiquement tout token comme une réponse potentielle. Un processus de filtrage basé sur la vraisemblance des prédictions est utilisé pour sélectionner les questions les plus pertinentes.

Nous proposons de baser notre processus de sélection des candidats *réponses* sur un processus d'analyse sémantique repérant dans un texte des relations sémantiques et de considérer chaque argument comme étant une réponse potentielle à une question portant sur la relation sémantique exprimée par le texte.

Le but de cette analyse sémantique est double : d'une part sélectionner des phrases susceptibles d'être pertinentes afin de servir de base à la génération de couples *question/réponse* car porteuses de l'expression d'une relation sémantique particulière ; et d'autre part de guider la génération de questions en modélisant explicitement le lien sémantique entre les réponses et les arguments des questions.

Deux types de représentation sémantique ont été testés dans cette étude : une représentation Berkeley FrameNet (Baker *et al.*, 1998), suivant les travaux précédents sur la génération de questions sur le corpus CALOR-QUEST (Béchet *et al.*, 2019), et le Semantic Role Labelling (SRL) tel qu'il a été proposé pour contrôler la génération de questions avec *BART* dans Pyatkin *et al.* (2021).

Afin d'entraîner le modèle de génération de question, nous utilisons le corpus FQuAD (d'Hoffschmidt *et al.*, 2020)) de questions/réponses en français collectées selon la même méthode que SQuAD. Nous appliquons la méthode suivante :

1. Annotation automatique avec FrameNet et les étiquettes SRL des contextes et des questions du corpus FQuAD. Seules les analyses déclenchées par un verbe sont gardées, l'hypothèse étant fait qu'elles portent une importance sémantique plus forte dans la phrase.
2. Pour chaque triplet question/réponse/contexte :
   (a) Trouver un rôle sémantique associé à la réponse d'une question donnée grâce à l'annotation effectuée. Pour cela, nous alignons les réponses et les rôles sémantiques, puis nous choisissons celui avec le recouvrement maximal.
   (b) Générer un exemple d'entraînement avec une séquence d'entrée concaténant le contexte, la réponse et éventuellement l'annotation sémantique. La question est ici la séquence de sortie.
3. Affinage d'un modèle de génération pré-entraîné *BART* sur le corpus ainsi obtenu.

Au moment de l'inférence, la génération de questions sur une phrase donnée consiste d'abord à effectuer une analyse sémantique de la phrase, puis à générer une séquence d'entrée pour chaque rôle sémantique détecté. Le modèle séquence-à-séquence affiné génère ensuite une question pour chacune d'entre elles. Dans cette étude, nous comparons 4 configurations différentes pour le format d'entrée du modèle séquence-à-séquence de génération de questions :

1. **basic-Frame-ctx** : dans cette représentation, seule la réponse extraite grâce au processus d'alignement avec les *Frame Elements* (*FrameNet*) comme décris en 2.(a) et le contexte sont donnés ;
2. **basic-SRL-ctx** : dans cette représentation, seule la réponse extraite grâce au processus d'alignement avec les roles sémantique (*PropBank*) comme décris en 2.(a) et le contexte sont donnés ;
3. **full-Frame-ctx** : dans cette représentation, le rôle de la réponse est rajouté et le contexte est également enrichi avec le déclencheur (ou `LU` ou *lexical unit*) de l'analyse *FrameNet*. Les autres *Frame Elements* de l'analyse sont également explicitement ajoutés ;
4. **full-SRL-ctx** : dans cette représentation, le rôle de la réponse est rajouté et le contexte est également enrichis avec les autres arguments de l'analyse *PropBank*. Le déclencheur n'est ici pas rajouté, correspondant simplement au lemme du verbe à l'origine de l'analyse sémantique.

L'exemple suivant, provenant de CALOR-QUEST, illustre ces 4 représentations à partir d'un contexte. Dans ce cas, les analyses *Frame Elements* et *PropBank* ont fourni le même ensemble de mots pour l'extraction de la réponse et les deux premières configurations sont ainsi identiques.

Context : Des outils du Paléolithique avec des dents et des os de mammouth ont été trouvés à [Flins-sur-Seine]answer
Question : Où a-t-on trouvé des outils du Paléolithique ?
basic-Frame-ctx : [ANS] Flins-sur-Seine [CTX] Des outils du Paléolithique avec des dents et des os de mammouth ont été trouvés à Flins-sur-Seine
basic-SRL-ctx : [ANS] Flins-sur-Seine [CTX] Des outils du Paléolithique avec des dents et des os de mammouth ont été trouvés à Flins-sur-Seine
full-Frame-ctx : [ANS:Location] Flins-sur-Seine [LU:Locating] trouvés [Sought-entity] des outils du Paléolithique [CTX] Des outils du Paléolithique avec des dents et des os de mammouth ont été trouvés à Flins-sur-Seine
full-SRL-ctx : [ANS:ARGM-LOC] Flins-sur-Seine [LU] trouvés [ARG1] des outils du Paléolithique [CTX] Des outils du Paléolithique avec des dents et des os de mammouth ont été trouvés à Flins-sur-Seine

3 Adaptation non supervisée d'un modèle de compréhension de documents

La tâche de compréhension de document consiste à localiser la réponse à une question donnée dans un texte. L'état de l'art consiste à affiner un modèle de langage pré-entraîné tel que BERT pour prédire les positions de début et de fin d'une réponse dans un paragraphe. La question et le paragraphe (contexte) sont donnés comme caractéristiques d'entrée.

Afin de réaliser cette tâche, nous avons utilisé un modèle de compréhension de lecture automatique basé sur un modèle de langage pour le français appelé CamemBERT (Martin *et al.*, 2020), affiné sur la tâche de question/réponse sur différents jeux de données comme il sera présenté dans la section 5. Le modèle de génération de questions sera utilisé pour produire, de manière non supervisée, le corpus d'entraînement nécessaire pour entrainer ou affiner un modèle de question/réponse sur du texte.

Dans cette étude, l'objectif est d'utiliser la génération automatique de questions afin de voir s'il est possible d'adapter de manière non supervisée un modèle de question/réponse à un autre corpus non annoté provenant d'un domaine différent de celui sur lequel le modèle a été appris initialement. En effet, même si les questions générées automatiquement seront, dans certains cas, imparfaites et qu'elles ne pourront pas correspondre à des questions de « difficultés » autres que celles rencontrées dans le corpus initial, nous faisons l'hypothèse que l'ajout de ces questions et contextes sur un nouveau domaine permettra de mieux prendre en compte de possibles différences lexicales et de structures dans ces nouveaux textes.

4 Le corpus *Autogestion*

L'autogestion couvre un grand nombre de notions du champ des sciences sociales. Elle concerne, l'environnement social quotidien, la vie économique, l'éducation, l'écologie, la culture, l'architecture... Elle évoque les structures de populations, le liens entre les populations et les ressources, l'organisation

politique, législative et administrative des sociétés, et les relations d'autorité entre les individus et les groupes. C'est une notion particulièrement transversale et interdisciplinaire qui peut alimenter des travaux de recherche en sociologie, sciences politiques, économie, droit, anthropologie politique et qui présente un intérêt tout particulier pour les chercheurs de la *Fondation Maison des Sciences de l'Homme* (FMSH [1]). Depuis les années 1960, la bibliothèque de la FMSH a constitué un fonds mixte (archives et documentation) sur l'autogestion, plurilingue et pluridisciplinaire. Il comprend environ 25 000 pièces : livres, revues, brochures, mémoires, rapports, tracts, comptes rendus de réunions, correspondances, disponibles suivant la charte CODHOS. Cette collection a reçu le label "Collection d'Excellence" par le réseau Collex-Persée [2].

Dans le but d'étudier la valorisation d'archives numérisées, nous travaillons plus particulièrement sur le corpus de la revue Autogestion [3]. Cette revue, constituée de 46 numéros publiés entre 1966 et 1986, constitue une ressource très riche pour les recherches en sciences sociales et l'accès à ce type de ressources via le paradigme du question/réponse représente un enjeu encore peu exploré. En effet, la majorité des travaux dans le domaine s'appuie sur des contenus issus de Wikipedia, où les textes essentiellement factuels se prêtent particulièrement bien à la tâche. Ici, les articles sont de nature variée, avec des articles reflétant des points de vue, des entretiens entre sociologues ou des études approfondies d'un courant de pensée ou d'un événement.

Nous utilisons une version numérisée avec Tesseract, pour un total de 6298 pages et 1,98 M de tokens. Le format de la revue est essentiellement en monocolonne avec peu de figures et même s'il serait important d'étudier l'impact de la qualité de l'OCR sur les tâches de TAL, nous n'abordons pas ce point ici et considérons la forme numérisée telle quelle.

Nous avons conduit une campagne d'annotation en questions de ce corpus, en proposant une annotation en différents types de questions. Les questions *faciles* sont proches des questions que l'on peut rencontrer dans les corpus de type SQuAD en ce sens qu'elles reprennent le champ lexical du paragraphe source et portent sur un empan bien identifié de ce paragraphe. Les questions dites *moyennes* cherchent à paraphraser plutôt qu'à reprendre les termes du texte d'origine. Les questions *difficiles* nécessitent une interprétation plus avancée du texte.

Exemples d'annotations :

- Question *facile* (niveau 1)
  - *Contexte* : `Pour l'heure, l'autogestion s'accompagne souvent d'une certaine étroitesse de vue.`
  - *Question* : **`De quoi l'autogestion s'accompagne-t-elle souvent ?`**
- Question *moyenne* (niveau 2)
  - *Contexte* : `Le marxisme a accouché tout aussi bien du réformisme que du léninisme et du stalinisme (sans parler du maoïsme).`
  - *Question* : **`Quelles ont été les conséquences du marxisme ?`**
- Question *difficile* (niveau 3)
  - *Contexte* : `Curieuse découverte en réalité que celle-ci. Aucun élément « scientifique » ne vient étayer la religion nouvelle de Bakounine !`
  - *Question* : **`À quoi est comparé le communisme ?`**

1. https ://www.fmsh.fr/
2. https://www.collexpersee.eu
3. https://www.persee.fr/collection/autog

5 Expériences

5.1 Modèles

Pour la génération de questions, le modèle BARThez dans sa version « base » (165M de paramètres) [4] (Eddine *et al.*, 2020) est affiné sur le corpus FQuAD, préalablement transformé avec les 4 variantes décrites à la section 2. L'affinage est effectué avec l'implémentation d'HuggingFace et une taille de batch de 1 ainsi qu'un nombre d'époques fixé à 5. Une évaluation sur les questions du jeu de validation de FQuAD est effectuée toutes les 500 étapes et le modèle retenu est celui obtenant le meilleur score BLEU sur celui-ci.

Pour la tache de compréhension de documents, nous affinons le modèle CamemBERT dans sa version « large » (335M de paramètres) [5]. Les affinages ont été réalisés avec des questions générées sur le corpus *Autogestion* à partir des quatre configurations décrites dans la section 2 ((**basic-Frame-ctx**, **basic-SRL-ctx**, **full-Frame-ctx**, et **full-SRL-ctx**). Les configurations se basant sur l'analyse de Frame de type FrameNet, en se restreignant aux déclencheurs verbaux, conduisent à un nombre de questions générées de 8513. Les configurations basées sur le SRL produisent plus de questions (49773) car l'analyse SRL est plus généraliste et une très grande majorité des occurrences de verbes génèrent une analyse en rôles sémantique. De façon à pouvoir comparer les approches avec un nombre identique de questions générées pour l'apprentissage des modèles de MRQA, nous avons extrait aléatoirement un ensmble de 8513 question à partir des configurations SRL. Ce modèle est appelé $M1$ dans nos expériences. Dans un second temps, un affinage du même modèle de base a été effectué en se basant sur le corpus d'entrainement de FQuAD (20731 exemples) enrichi de 1000 exemples. Les 1000 exemples sont tirés aléatoirement parmi les 8513 questions générées, le nombre de 1000 a été choisi empiriquement [6] et représente approximativement 5% de la taille de $FQuAD_{train}$. L'affinage est effectué avec l'implémentation d'HuggingFace et une taille de batch de 6 ainsi qu'un nombre d'époques fixé à 4. Une évaluation sur le jeu de validation de FQuAD est effectuée toutes les 1000 étapes et le modèle retenu est celui obtenant le meilleur score sur celui-ci. Ce modèle est appelé $M2$ dans nos expériences.

Dans la première configuration ($M1$), le modèle est appris exclusivement sur des questions synthétiques, simulant ainsi les performances pouvant être obtenues sur un nouveau type de texte issu d'un domaine spécifique, sans aucune annotation. Dans la seconde ($M2$), les questions synthétiques sont utilisées pour adapter un modèle appris sur des données annotées sur un corpus généraliste issu de Wikipedia (le corpus FQuAD).

5.2 Résultats

Tous les résultats qui suivent sont la moyenne et l'écart type sur trois modèles à graines aléatoires fixes (2, 7 et 9). Les modèles sont évalués sur le corpus Autogestion (842 questions sur 3 niveaux de difficultés) en termes d'Exact-Match, micro-F1, précision et rappel.

Dans un premier temps, nous évaluons dans la table 1 comment se comporte le modèle $M1$ sur le corpus Autogestion. Ceci permettra de déterminer si la génération automatique seule permet

4. https://huggingface.co/moussaKam/barthez
5. https://huggingface.co/camembert/camembert-large
6. ce choix est discuté à la section ? ?

	Niveau 1 (416 questions)				Niveau 2 (352 questions)				Niveau 3 (74 questions)			
Modèle M1	EM	F1	Prec.	Rappel	EM	F1	Prec.	Rappel	EM	F1	Prec.	Rappel
basic-Frame-ctx	20.5 (±1.8)	36.7 (±1.4)	45.3 (±1.9)	30.9 (±2.3)	9.1 (±0.8)	25.4 (±1.9)	32.8 (±0.6)	20.8 (±2.5)	1.3 (±0.0)	17.5 (±1.1)	19.1 (±0.4)	16.4 (±2.0)
basic-SRL-ctx	**24.8** (±0.3)	**43.4** (±1.5)	53.6 (±2.4)	**36.7** (±3.0)	**13.2** (±0.6)	**31.8** (±1.3)	**41.6** (±1.9)	**26.0** (±2.5)	**5.4** (±2.2)	18.4 (±1.6)	20.9 (±1.9)	16.6 (±1.9)
full-Frame-ctx	21.0 (±0.8)	39.9 (±0.2)	48.0 (±1.4)	34.2 (±0.9)	9.7 (±0.3)	27.7 (±1.9)	34.7 (±1.7)	232 (±2.1)	1.3 (±1.1)	20.7 (±1.7)	22.5 (±0.9)	**19.2** (±2.2)
full-SRL-ctx	24.3 (±1.4)	41.9 (±2.8)	**54.9** (±2.1)	33.9 (±2.9)	11.8 (±1.0)	29.2 (±2.6)	39.7 (±2.9)	23.1 (±2.2)	3.1 (±0.6)	**21.2** (±0.9)	**24.6** (±2.1)	18.67 (±0.4)

TABLE 1 – Performances du modèle $M1$ sur la tâche de question/réponse selon le niveau de difficulté des questions du corpus

	Niveau 1 (416 questions)				Niveau 2 (352 questions)				Niveau 3 (74 questions)			
Modèle M2	EM	F1	Prec.	Rappel	EM	F1	Prec.	Rappel	EM	F1	Prec.	Rappel
FQuAD seul	37.5 (±0.7)	62.9 (±1.0)	80.9 (±0.4)	51.4 (±1.1)	25.1 (±0.3)	**54.8** (±0.5)	**69.8** (±0.7)	**45.1** (±0.4)	18.0 (±0.6)	**48.2** (±3.7)	**58.2** (±6.0)	**41.1** (±2.6)
FQuAD + basic-Frame-ctx	39.9 (±1.4)	62.8 (±1.3)	80.1 (±1.8)	51.7 (±1.0)	27.0 (±1.4)	53.5 (±0.9)	67.6 (±1.7)	44.3 (±0.8)	16.2 (±1.1)	43.0 (±0.4)	47.7 (±1.0)	39.2 (±0.8)
FQuAD + basic-SRL-ctx	**40.9** (±1.6)	**64.2** (±0.1)	80.8 (±0.6)	**53.3** (±0.1)	26.8 (±2.7)	53.1 (±0.7)	66.8 (±0.8)	44.0 (±0.7)	**19.8** (±2.3)	44.9 (±2.4)	49.9 (±2.3)	40.8 (±2.7)
FQuAD + full-Frame-ctx	40.6 (±0.9)	63.4 (±0.4)	80.1 (±0.6)	52.5 (±0.5)	**27.37** (±1.6)	53.8 (±0.7)	67.1 (±0.3)	44.9 (±1.05)	19.4 (±0.6)	44.3 (±1.7)	48.5 (±0.7)	40.8 (±3.3)
FQuAD + full-SRL-ctx	39.9 (±1.6)	62.8 (±1.6)	**81.4** (±1.9)	51.1 (±1.4)	26.9 (±1.1)	52.6 (±1.1)	68.3 (±2.0)	42.8 (±1.4)	16.7 (±3.5)	41.4 (±2.3)	48.4 (±2.4)	36.5 (±4.2)

TABLE 2 – Performance du modèle $M2$ appris sur FQuAD et adapté sur le corpus *Autogestion*

d'apprendre des modèles sur un nouveau domaine. On constate que les résultats en termes de correspondance exacte (*Exact Match - EM*) et de *F-mesure* par rapport aux réponses de référence sont relativement bas, y compris dans le cas de questions de niveau 1 avec des scores approchant les 25% d'EM et 44% de F1. La qualité insuffisante des couples questions-réponses générés automatiquement est probablement en cause ici. On remarque aussi que l'ajout de détails de l'analyse sémantique lors de la génération de question (**full-Frame-ctx** et **full-SRL-ctx**) n'améliore pas les résultats.

Dans un second temps, nous évaluons le modèle $M2$ appris sur le corpus FQuAD auquel $1\,000$ questions générées automatiquement sur le corpus Autogestion sont ajoutées. La table 2 présente les résultats de l'évaluation. Afin d'avoir un point de comparaison, les résultats obtenus lorsque le modèle est uniquement appris sur FQuAD est également présenté.

Si on ne s'intéresse qu'aux questions de niveau 1, on constate un gain en EM (37.5 vs 40.9) et F1 (62.9 vs 64.2) lors de l'ajout de questions de type **SRL**. Ce gain ne se retrouve pas avec l'ajout des questions de type **Framenet**, ce qui pourrait être dû à la plus grande précision de l'annotation FrameNet, possiblement plus difficile à correctement exploiter pour un modèle de génération de questions. Sur les questions de niveaux 2 et 3, l'intérêt de l'ajout de questions automatiques n'est pas clair. En effet, les scores F1 sont inférieurs au modèle appris uniquement sur FQuAD.

Néanmoins, il est intéressant de constater que les questions de type **Framenet** permettent d'obtenir des scores de F1 inférieur (niveau 3 : 44.9 vs 48.2), mais plus important en EM (niveau 3 : 18.0 vs 19.8).

Enfin, de manière générale, on peut remarquer que seules les questions de niveau 1 affichent des performances convenables, les questions plus abstraites de niveau 2 et 3 obtiennent des résultats

jusqu'à 3 fois inférieures à ce qui peut être obtenus avec les meilleurs modèles état-de-l'art sur des corpus comme SQuAD.

	Résultats globaux			
Modèle M1	EM	F1	Prec.	Rappel
basic-Frame-ctx	14.0 (±1.2)	30.2 (±1.5)	37.7 (±1.2)	25.4 (±2.4)
basic-SRL-ctx	**18.2** (±0.1)	**36.5** (±1.3)	**45.9** (±2.0)	**30.5** (±2.7)
full-Frame-ctx	14.6 (±0.4)	33.1 (±1.0)	40.2 (±0.8)	28.2 (±1.5)
full-SRL-ctx	17.3 (±1.0)	34.8 (±2.5)	**45.9** (±2.4)	28.0 (±2.4)

	Résultats globaux			
Modèle M2	EM	F1	Prec.	Rappel
FQuAD seul	30.5 (±0.3)	**58.2** (±0.7)	**74.2** (±0.9)	47.9 (±0.6)
FQuAD + basic-Frame-ctx	32.3 (±1.0)	57.2 (±1.0)	71.8 (±1.7)	47.6 (±0.8)
FQuAD + basic-SRL-ctx	33.1 (±2.0)	57.9 (±0.2)	72.1 (±0.2)	**48.3** (±0.3)
FQuAD + full-Frame-ctx	**33.2** (±1.0)	57.7 (±0.2)	71.7 (±0.6)	**48.3** (±0.6)
FQuAD + full-SRL-ctx	32.4 (±1.1)	56.7 (±1.2)	72.9 (±1.3)	46.5 (±1.4)

TABLE 3 – Performances globales des modèles $M1$ et $M2$ sur la tâche de question/réponse

6 Discussions et analyses

Dans cette section, nous menons quelques expériences contrastives et nous discutons les résultats obtenus, et ce pour la tâche de compréhension de lecture (MRQA)

6.1 Impact de l'ajout de questions synthétiques à FQuAD

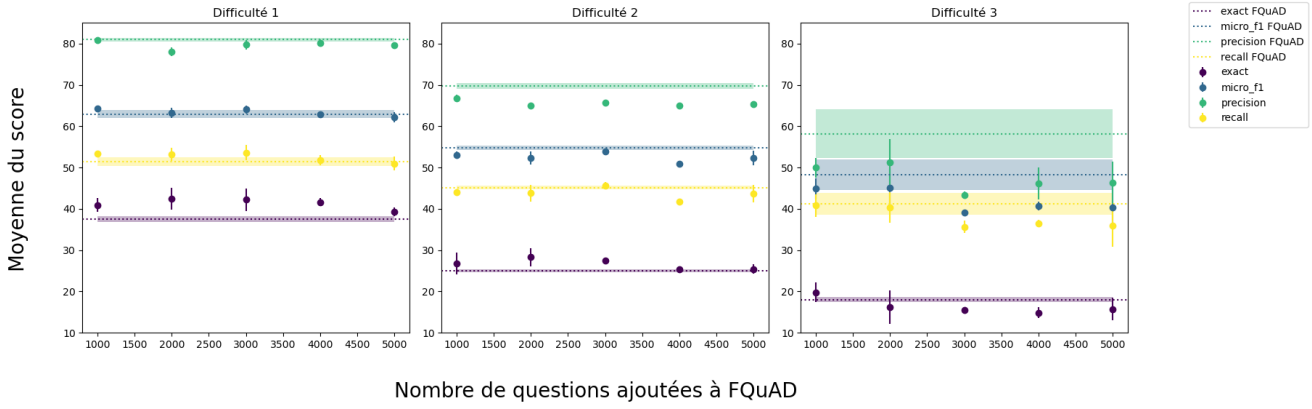


FIGURE 1 – Scores en fonction du nombre d'exemples de la configuration **basic-SRL-ctx** ajoutés.

Pour mesurer l'impact du nombre de questions synthétiques rajoutées à FQuAD nous avons ici entraîné d'autres instances du modèle *M2* dans les mêmes conditions, en faisant uniquement varier le nombre de questions synthétiques.

La figure 1 présente les résultats de ces modèles en fonctions du nombre de questions ajoutées sur le même corpus d'évaluation que dans la table 2. Nous comparons également ici les résultats de ces modèles à ceux obtenus avec un modèle appris uniquement sur FQuAD (ligne en pointillé pour la moyenne et boite pour l'écart type associé). On peut remarquer que les meilleures performances sont

Question	Que reste-t-il de Marx ?
Difficulté	3
Réponse de référence	l'autoémancipation ouvrière ne peut être que sociale et le moyen n'en est pas la conquête et la transformation de l'Etat, mais l'abandon et la destruction de tout pouvoir politique
Réponse prédite avec basic-SRL-ctx	l'autoémancipation ouvrière ne peut être que sociale
Réponse prédite avec FQuAD + basic-SRL-ctx	un acquis
Contexte	[...] La première, la « conquête de la démocratie » par la classe ouvrière, débouche sur la « dictature du prolétariat ». La seconde, c'est l'abolition des classes sociales et du pouvoir politique, la naissance de la société humaine » (p. 177) Ainsi, « pour ambigü qu'ait été l'héritage de Marx, il en reste pourtant un acquis : l'autoémancipation ouvrière ne peut être que sociale et le moyen n'en est pas la conquête et la transformation de l'Etat, mais l'abandon et la destruction de tout pouvoir politique » (p. 213). [...]

TABLE 4 – Exemple difficile (**3**) de prédictions pour le MRQA sur plusieurs configurations

obtenus lors de l'ajout d'un nombre relativement faible de questions en comparaison à la taille du jeu de données de base (1000 ou 2000 questions en fonctions de la difficulté contre environ 20 000 pour FQuAD). On peut supposer que cela est dû au fait que l'ajout d'un trop grand nombre d'exemples synthétiques bruite trop fortement les exemples de référence présents dans le jeu de données. Cela a alors pour conséquence, en plus d'annuler les gains obtenus principalement en difficulté 1 et 2 pour l'exact-match et la micro-F1, de baisser encore plus les performances obtenues sur les autres scores et difficultés.
La figure 1 montre égalament que la dégardation des performances obtenue lorsque l'on ajoute trop de données d'apprentissage synthétique peut s'observer quel que soit le niveau de difficulté, et quelle que soit la métrique. En revanche la dégradation est plus importante pour les questions difficiles. De plus, pour les questions considérées les plus difficiles (difficulté **3**) on observe un écart type sur les performances inter-modèles bien plus élevé, allant jusqu'à 2.9 points pour la précision. Il est néanmoins assez difficile ici d'expliquer cette variation inter-modèle uniquement par une difficulté plus élevée, la faible quantité d'exemples dont nous disposons pour celle-ci (**74**) pouvant tout aussi bien en être la cause.

En conclusion, dans un paradigme d'adaptation par augmentation de données avec des questions générées automatiquement, l'ajout doit être limité. Une piste de travail sera de sélectionner de façon plus appropriée que le tirage aléatoire ces 1000 questions à ajouter au corpus d'apprentissage.

6.2 Discussion sur l'adéquation du score F1 avec la pertinence réelle des réponses prédites

L'empan des réponses de référence et les choix adoptés par les annotateurs pour délimiter ces réponses, ont un impact important sur l'annotation. Dans le cadre du corpus FQuAD, ou les questions sont majoritairement factuelles sur des données encyclopédiques, les réponses sont généralement assez courtes, et annotées sans prépositions ou déterminants en début de segment de réponse (4,2 mots en moyenne dans les réponses FQuAD). Le corpus *Autogestion* étant plus complexe et les questions posées pouvant être de niveau de difficulté variable, les réponses sont globalement plus longues (9,2 mots en moyenne par réponse). De ce fait le corpus *Autogestion* est plus impacté par la métrique F1 qui compare les empans des réponses.

Les tables 4 et 5 présentent des exemples de prédictions pour la tâche de MRQA sur le corpus *Autogestion*. La réponse produite par la configuration (**basic-SRL-ctx**) dans le cas du modèle *M1* (questions générées seules) et *M2* (questions générées concaténées à FQuAD) est comparée à la réponse de référence. Seule la partie pertinente du contexte est donnée ici.

On peut observer dans la table 4 un exemple de question considéré comme difficile. Celui-ci est

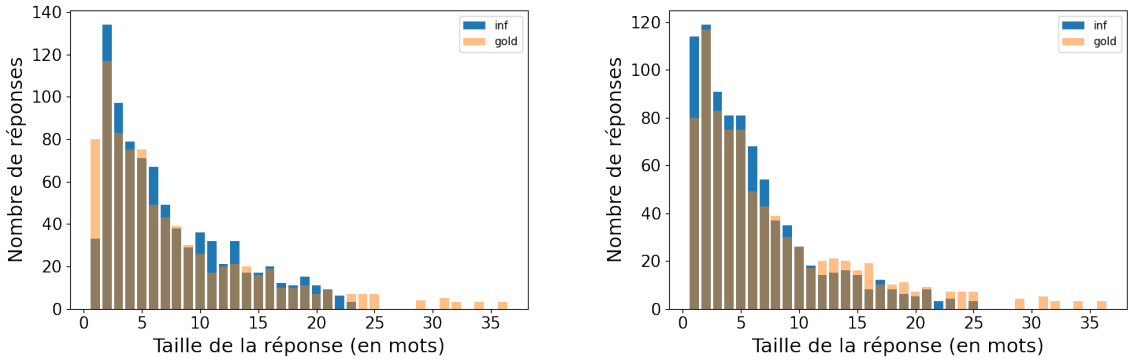
Question	**De quelle caste se compose l'État ?**
Difficulté	1
Réponse de référence	la caste des « hauts fonctionnaires »,
Réponse prédite avec basic-SRL-ctx	la caste des « hauts fonctionnaires
Réponse prédite avec FQuAD + basic-SRL-ctx	hauts fonctionnaires
Contexte	[...] « Centralisation hiérarchique et unique des pou- voirs publics » l'Etat, ou plutôt la caste des « hauts fonctionnaires », qui animent son appareil se constitue, sur l'antagonisme des pouvoirs sociaux, et l'entretient. [...]

TABLE 5 – Exemple facile (**1**) de prédictions pour le MRQA sur plusieurs configurations

particulièrement intéressant, car il représente les cas où l'utilisation de FQuAD en plus des questions générées n'améliore pas la prédiction, la dégradant au contraire fortement. En effet le modèle $M2$ prédit une réponse très courte et non informative. Le modèle $M1$ produit une réponse pertinente, mais le score F1 sera néanmoins très bas du fait du rappel pénalisé par la longuer de la répone de référence.

Dans d'autres cas de figures, illustrés dans la Table 5, une prédiction "correcte" d'un point de vue sémantique, mais mal alignée sur le niveau de détail choisi par l'annotateur, conduira à un F1-score bas. Cela arrive assez souvent dans notre cas, particulièrement dans le cas du modèle *M2*, où l'apprentissage se fait sur FQuAD. La majorité des exemples du modèle *M2* venant de FQuAD, il est alors encouragé à prédire des réponses plus courtes, correspondant à celles vues lors de son apprentissage (figures 2a et 2b).

Ces deux exemples sont en défaveur du modèle $M2$, mais bien que les réponses soient en moyenne plus courtes avec ce système, celui-ci "cible" mieux la prédiction et obtient donc globalement de meilleurs scores que le modèle *M1*. L'objectif ici était de montrer le comportement de la métrique F1.



(a) basic-Frame-ctx

(b) FQuAD + basic-SRL-ctx

FIGURE 2 – Répartition des réponses prédites (**inf**) en fonction de leur longueur et comparaison avec les réponses de référence (**gold**) sur le jeu de test *Autogestion*

En conclusion, l'évaluation de la tâche de compréhension de lecture sur des textes plus complexes que les textes encyclopédiques et pour des questions plus complexes que les simples questions factuelles, nécessiterai de nouvelles métriques tenant compte de la pertinence des réponses produites.

7 Conclusion

Nous avons présenté dans cette étude une méthode d'adaptation automatique de modèles de compréhension de document basée sur la génération de question à partir d'analyse sémantique. Les résultats obtenus sur un corpus *difficile* permettent d'une part de montrer l'intérêt d'une telle approche pour améliorer les résultats obtenus avec un modèle *généraliste* appris sur Wikipedia (FQuAD) et d'autre part de mesurer l'effort qu'il reste à faire pour utiliser ces modèles sur des questions plus réalistes d'un point de vue applicatif que les questions littérales simples des corpus de références tels que SQuAD.

Remerciements

Ces travaux ont été partiellement financés par l'Agence Nationale pour la Recherche (ANR) à travers le projet ANR-19-CE38-0011 (ARCHIVAL).

Ces travaux ont bénéficié d'un accès aux ressources en HPC/IA de l'IDRIS au travers de l'allocation de ressources 2021-AD011012688 attribuée par GENCI.

Ces travaux ont bénéficié d'une aide du gouvernement français au titre du Programme Investissements d'Avenir Initiative d'Excellence d'Aix-Marseille Université - A*MIDEX (Institut Archimède AMX-19-IET-009).

Références

Baker C. F., Fillmore C. J. & 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*, p. 86–90 : Association for Computational Linguistics.

Béchet F., Aloui C., Charlet D., Damnati G., Heinecke J., Nasr A. & 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*.

d'Hoffschmidt M., Belblidia W., Heinrich Q., Brendlé T. & Vidal M. (2020). FQuAD : French question answering dataset. In *Findings of the Association for Computational Linguistics : EMNLP 2020*, p. 1193–1208, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.findings-emnlp.107.

Du X., Shao J. & 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)*, p. 1342–1352.

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

Martin L., Muller B., Ortiz Suárez P. J., Dupont Y., Romary L., de la Clergerie É., Seddah D. & Sagot B. (2020). CamemBERT : a tasty French language model. In *Proceedings of*

the 58th Annual Meeting of the Association for Computational Linguistics, p. 7203–7219, Online : Association for Computational Linguistics.

PURI R., SPRING R., SHOEYBI M., PATWARY M. & CATANZARO B. (2020). Training question answering models from synthetic data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 5811–5826.

PYATKIN V., ROIT P., MICHAEL J., GOLDBERG Y., TSARFATY R. & 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*, p. 1429–1441, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.108.

RAJPURKAR P., ZHANG J., LOPYREV K. & 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*, p. 2383–2392 : Association for Computational Linguistics. DOI : 10.18653/v1/D16-1264.

SHAKERI S., DOS SANTOS C., ZHU H., NG P., NAN F., WANG Z., NALLAPATI R. & 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)*, p. 5445–5460.