Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Questionner pour expliquer : construction de liens explicites entre documents par la génération automatique de questions

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

RÉSUMÉ

Cette article présente une méthode d'exploration de documents basée sur la création d'un ensemble synthétique de questions et de réponses qui est ensuite utilisé pour établir des liens explicables entre les documents. Nous menons une évaluation quantitative et qualitative des questions automatiquement générées en termes de leur forme et de leur pertinence pour l'exploration de la collection. De plus, nous présentons une étude quantitative des liens obtenus grâce à notre méthode sur une collection de document provenant d'archives numérisés.

ABSTRACT

Expliciting links between documents through automatic question generation

This paper presents a method for document mining based on the creation of a synthetic set of questions and answers. This set is used to establish explainable links between documents. We conduct a quantitative and qualitative evaluation of the automatically generated questions in terms of their form and their relevance to the exploration of the collection. In addition, we present a quantitative study of the links obtained with our method on a collection of documents from a digitized archive.

MOTS-CLÉS : Génération de questions ; Compréhension de documents ; Explicabilité.

KEYWORDS: natural language understanding ; question generation ; explainability.

1 Introduction

Cette étude porte sur l'exploration de collections de documents par des méthodes d'appariements permettant de créer des liens entre des textes de la collection. De nombreuses méthodes basées sur des mesures de similarité permettent de créer de tels liens, cependant la justification des liens pour les utilisateurs est bien souvent réduite à la présence de mots clés en commun. Nous proposons dans cette étude d'utiliser le paradigme des questions/réponses sur des textes afin de créer des liens explicables : deux paragraphes sont liés si les questions que posent ces paragraphes sont également liés.

Cet article est structuré comme suit : la section 2 présente notre méthodologie pour générer des liens explicables entre les documents basés sur des modèles de génération de questions ; la section 3 présente notre méthode de génération et de filtrage de questions ; la section 4 décrit comment les questions générées peuvent être utilisées pour créer des liens explicables entre les documents ; enfin,

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 4 : articles déjà soumis ou acceptés en conférence internationale, pages 1–9.

les sections6 et 7 présentent une étude expérimentale sur notre corpus d'archives avec des évaluations quantitatives, descriptives et qualitatives de la méthode proposée.

2 Exploration par la génération de questions

Lors de l'exploration d'une collection d'archives thématiques, des liens peuvent être établis entre des documents ou des parties de documents en fonction de différents critères tels que la co-occurrence d'entités (personnes, lieux, organisations, dates, ...), de mots clés liés à une base de connaissances ou à un thésaurus (Tsatsaronis *et al.*, 2014), ou directement par une mesure de similarité statistique entre des documents ou des parties de documents tels que des phrases (Wang *et al.*, 2016) ou des paragraphes (Dai *et al.*, 2015). La structure en graphe ainsi obtenue peut être utilisée pour concevoir des interfaces de navigation telles que des cartographies ou directement en insérant des liens hypertextes.

Générer des liens basés sur les mots clés/entités pose plusieurs problèmes : d'une part la grande quantité de liens générés si de grands ensembles de mots clés ou d'entités sont considérés et d'autre part le fait que la simple occurrence de termes pertinents ne signifie pas que leurs contextes d'occurrence sont pertinents ou intéressants pour les utilisateurs.

Les liens basés sur la similarité permettent de corriger ce dernier point en prenant en compte les mots dans leur contexte, mais l'utilisation de mesures de similarité statistique rend souvent les liens difficiles à interpréter, obligeant les utilisateurs à les vérifier un par un pour évaluer leur pertinence, ce qui peut être très chronophage.

Récemment, les avancées dans les modèles de Question-Réponse (QR) à partir de texte ont permis l'utilisation de questions directes en langage naturel afin d'accéder à des documents électroniques. Des résultats impressionnants ont été obtenus avec des modèles de langage pré-entraînés tels que BERT sur des corpus de référence comme SQuAD (Rajpurkar *et al.*, 2016). Cependant il a été montré que le type de questions que ces modèles gèrent le mieux sont les questions littérales simples pour lesquelles une réponse factuelle peut être trouvée dans le texte en une seule phrase, et que les performances diminuent lorsque l'on traite de questions plus abstraites ou nécessitant un contexte plus large qu'une phrase pour être abordées. De plus, la plupart de ces études ont été appliquées uniquement au texte de Wikipedia.

Dans notre étude nous n'allons pas utiliser le mode *question→réponse* pour accéder au texte, mais plutôt le mode *réponse→question* : en sélectionnant des zones potentiellement *intéressantes* dans nos corpus nous allons générer des questions à partir du contenu de ces zones (les *réponses*), puis chercher des similarités entre les questions générées pour proposer des liens aux utilisateurs qui soient motivés par la paire de questions mises en relation et qui constitue l'*explication* du lien pour l'utilisateur.

3 Génération de questions

Pour la génération de question, nous utilisons une variation d'un des modèles proposés par (Bechet *et al.*, 2022), où une annotation sémantique (SRL) suivant le formalisme PropBank (Palmer *et al.*, 2005) est effectuée afin de sélectionner des *réponses* potentielles pour générer des questions. Cette sélection par un analyseur sémantique nous permet de ne garder que des contextes potentiellement riches

en terme de sens qui peuvent servir de supports à l'expression d'un lien avec d'autres documents.

La génération de question est vue comme une tâche de génération de texte de type *séquence-à-séquence* avec le modèle *BARThez* (Kamal Eddine *et al.*, 2021) entraîné sur le corpus FQuAD (d'Hoffschmidt *et al.*, 2020) de questions/réponses en français. La séquence en entrée contient la *réponse* (ANS), l'unité lexicale qui déclenche la relation sémantique LU et le contexte CTX, comme dans l'exemple suivant provenant de notre jeu d'entrainement *FQuAD* :
source : `[ANS:ARG2] Héra [LU] appelée[CTX] Cérès fut également appelée Héra en Allemagne pendant une brève période.`
cible : `Quel nom Cérès a-t-elle porté pendant une brève période en Allemagne ?`

Nous appliquons une série de filtres pour améliorer la qualité et réduire la quantité d'exemples générés. La première étape (**F1**) consiste à restreindre l'analyse SRL pour n'inclure que les cadres ayant un déclencheur strictement verbal (en rejetant les verbes auxiliaires), car ceux-ci sont considérés comme étant de meilleure qualité en raison de leur facilité de détection.

Pour améliorer encore la qualité des exemples générés, nous appliquons un filtre (**F2**) sur les requêtes afin d'éliminer celles dont les réponses ou les contextes ne sont pas informatifs. Cela inclut les réponses de moins de 5 caractères, ou appartenant à la liste NLTK (Bird *et al.*, 2009) des mots d'arrêt, afin d'éliminer les réponses contenant uniquement des coréférences pronominales. Les requêtes avec un contexte de moins de 5 mots ou une réponse qui n'est pas située dans la phrase du déclencheur sont également éliminées.

Les questions générées sont également soumises à un filtre (**F3**) basé sur la méthodologie "roundtrip consistency" proposée par (Alberti *et al.*, 2019). Ce filtre consiste à ne conserver que les exemples synthétiques pour lesquels un modèle de question-réponse [1] est capable d'extraire une partie de la réponse cible de la question générée. Nous considérons que le modèle a réussi à récupérer la réponse s'il y a un chevauchement minimum de 30 % entre la réponse prédite et la réponse de la requête.

Enfin, nous appliquons un dernier filtre (**F4**) pour éliminer les questions dupliquées, qui sont un phénomène fréquent dû à de légères variations dans certaines requêtes, résultant souvent en des questions très similaires ou identiques.

4 Générer des liens explicables

L'originalité principale de notre approche est l'utilisation de nos questions/réponses synthétiques pour établir des liens entre les documents de notre corpus. Alors que les méthodes traditionnelles consistent à calculer la similarité via les plongements de documents à un niveau de granularité choisi (phrase, paragraphe ou bloc de texte, page), notre approche consiste à calculer une mesure de similarité sur des plongements obtenus par la concaténation de questions et de réponses produites par notre méthode décrite dans la section 3.

Nous obtenons des structures "`<question> | <réponse>`". Par exemple, voici la structure obtenue sur l'exemple de la question générée donnée dans la section précédente : **`Quel nom Cérès a-t-elle porté pendant une brève période en Allemagne ? | Héra`**

Notre plongement pour chaque paire question-réponse utilise la bibliothèque SentenceTransformer

1. Dans notre cas *CamemBert-large* (Martin *et al.*, 2020) entrainé sur *FQuAD*

(Reimers & Gurevych, 2019) [2]. Une mesure de similarité cosinus est ensuite utilisée entre toutes les paires de ces projections, ce qui donne lieu au calcul d'une matrice de similarité.

Les sections suivantes présentent une étude expérimentale sur un corpus d'archives numérisées provenant du domaine des sciences sociales.

5 Thèmes des questions

Nous définissons les thèmes de la question générée par rapport à un thésaurus spécifique qui a été créé pour le domaine de l'autogestion. À partir de connaissances préalables dans le domaine, une première liste de notions a été construite. Elle a ensuite été enrichie par une liste de mots-clés et de phrases-clés extraites des articles de la revue "Autogestion". Ces termes sont principalement des phrases nominales extraites grâce à une analyse morphosyntaxique des documents. Lorsque des variantes flexionnelles d'une locution sont rencontrées, la forme ayant le plus grand nombre d'occurrences est choisie (forme majoritaire). Parmi toutes les phrases-clés extraites, les experts ont sélectionné une liste d'entrées supplémentaires pour le thésaurus en choisissant des termes qui font référence à des notions générales pouvant être pertinentes pour indexer des documents.

Le thésaurus est ensuite trié hiérarchiquement pour former une structure en arbre de profondeur maximale de quatre niveaux. L'arbre a 437 feuilles et est organisé en huit notions générales à la racine de l'arbre (*Organisations, Classes sociales, Développement économique, Exercice du pouvoir, Juridique, Modèles politiques, Psycho-sociologie, Valeurs sociales*).

$\lvert w \in T \rvert$	0	1	2	3	4	5
$\lvert Q + R \rvert$	43693	25159	8472	2129	355	58

TABLE 1 – Répartition du nombre de mots (w) appartenant au thésaurus T parmi les questions+réponses ($Q + R$)

terme	**occurrences**
travailleurs	3247
travail	2668
pouvoir	2214
société	1947
révolution	1834
production	1742
contrôle	1470
système	1426
ouvrier	1387
mouvement	1362

TABLE 2 – Les dix termes les plus fréquents du thésaurus

Nous avons analysé notre corpus de paires de questions/réponses synthétiques pour étudier l'utilisation des entrées du thésaurus. Nos résultats montrent que **30,6%** des questions générées et **25,1%** des

2. Nous utilisons le modèle multilingue *distiluse-base-multilingual-cased-v1* (Reimers & Gurevych, 2020)

réponses contiennent au moins un terme du thésaurus. De plus, nous avons constaté que **45,3%** des paires question/réponse incluent au moins un mot du thésaurus dans la question ou dans la réponse, et **10,4%** contiennent un mot du thésaurus dans les deux. Une description plus détaillée de la répartition du nombre d'entrées détectées dans les questions et les réponses peut être trouvée dans le Tableau 1 et les 10 entrées les plus fréquentes dans le Tableau 2.

La paire avec le plus de termes du thésaurus est la suivante : `Q : Qu'est ce qui rendra possible le` **`développement`** `de la` **`participation`** `des` **`travailleurs`** `et de leurs` **`organisations`** `à la direction et à la gestion des entreprises nationales ? A : le` **`changement`** `-- en` **`droit`** `et dans les faits -- des formes de la` **`propriété`**

Cette analyse suggère qu'en plus de permettre la création de liens pour explorer la collection, les questions générées pourraient également être un moyen d'illustrer les principales notions abordées dans le journal. Des interfaces dédiées pourraient être développées à cette fin dans le cadre de travaux futurs.

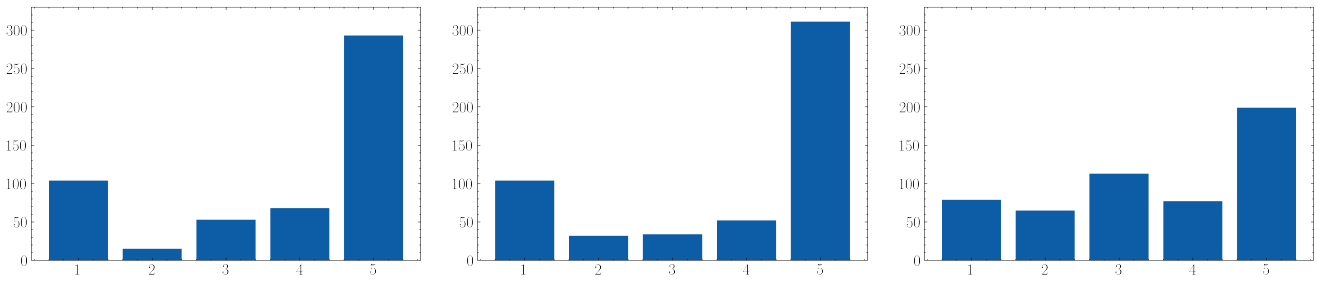
6 Évaluation qualitative des questions générées

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

TABLE 3 – Résumé des différents filtres : F2 (suppression des réponses non informatives), F3 (*round-trip consistency*), F4 (suppression des doublons).

Nous avons appliqué notre méthode de génération de questions à un corpus de 24 numéros de la collection *Autogestion* provenant d'un fond d'archives en sciences sociales allant de 1966 à 1979. Chaque numéro contient plusieurs articles courts ou longs pour un total de 448 articles. La version électronique de ce corpus étant obtenue par OCR, nous disposons de deux niveaux de segmentation supplémentaires : *page* (correspondant à l'OCR de chaque image d'une page donnée de la collection) et *bloc de texte* (l'unité minimale de texte cohérent produite par le système OCR). Nous considérons ici un sous-ensemble du corpus entier, contenant 4786 pages, 33551 blocs de texte pour un total de 1,5 million de tokens. Initialement, le processus d'étiquetage des rôles sémantiques produit 143 317 détections de cadres, qui sont réduites à 124 925 lorsque l'on se concentre sur les verbes non auxiliaires à partir du processus de filtrage **F1**. Chaque détection de cadre sémantique produit en moyenne 1,7 élément de trame, ce qui signifie que le premier ensemble est composé de 247 907 questions. Le tableau 3 indique le nombre de questions générées à la suite des processus de filtrage décrits dans la Sextion 3.

En plus de l'évaluation quantitative et descriptive des questions générées sur le corpus de *self-management*, nous avons également effectué une première évaluation qualitative sur un sous-ensemble de la collection selon deux dimensions. La première dimension se concentre sur la qualité de la forme de la question, avec des questions catégorisées comme "*Valide*", "*Question incohérente*" ou "*Question non grammaticale*". Dans la deuxième dimension, nous évaluons la pertinence de la question une fois qu'elle a été validée dans la dimension précédente. Cette évaluation implique trois échelles de Likert à 5 points :



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

FIGURE 1 – Évaluation de la pertinence des questions générées

1. "Le segment mis en évidence correspond bien à une réponse à la question"
2. "La question est pertinente dans le contexte de la phrase"
3. "La question est pertinente dans le contexte général de la lecture"

Un annotateur professionnel a été embauché pour cette tâche et a annoté un total de 582 questions. Environ 92% des questions ont été validées sur leur forme, ce qui confirme la qualité syntaxique de notre système de génération de questions.

Pour les annotations de pertinence, les résultats sont également prometteurs. En termes d'adéquation de réponse, la majorité des questions (67%) ont reçu une note indiquant un niveau élevé d'adéquation [3] (Figure 1a). Les deux échelles de Likert mesurant la pertinence de la question sont plus subjectives, mais une grande proportion de questions (plus de 68%) ont été évaluées comme pertinentes dans le contexte local (Figure 1b). Dans le contexte global de la lecture, le pourcentage de questions évaluées comme pertinentes diminue, avec un peu plus de la moitié des questions répondant au même critère de score (Figure 1c).

Pour vérifier l'accord entre les annotateurs, un sous-ensemble de 129 questions a été annoté par un second annotateur. Pour la première dimension (forme de surface), nous avons remarqué seulement 11 désaccords entre les deux annotateurs. Pour la seconde, concernant le premier Likert avec une évaluation simplifiée en 3 catégories en regroupant les choix 1 et 2 et les choix 4 et 5, nous avons mesuré 25 désaccords sur 115 annotations. Avec le même regroupement, nous avons obtenu 43 désaccords sur 115 annotations pour le Likert 2 et 65 désaccords sur 115 annotations pour le Likert 3. Ces nombres plus élevés de désaccords étaient attendus, car cette dernière évaluation est hautement subjective.

7 Évaluation des liens entre documents

L'évaluation des liens produits grâce aux questions nécessite une évaluation auprès d'utilisateurs dans un contexte réaliste étant donné la forte subjectivité de cette évaluation. Une évaluation de ce type est en cours. Cependant il est également possible d'évaluer ces liens par des mesures complémentaires automatiques. Ainsi nous avons évalué notre approche d'appariement basée sur les questions par rapport à des méthodes d'appariement uniquement basées sur la similarité textuelle. Pour quantifier la

3. Dans ce paragraphe, la notion de haut niveau d'adéquation correspond aux scores Likert > 3

	Pourcentage d'intersection		
Ensembles de similarités	(ALL)	(OUT_ART)	(OUT_NUM)
[QA] // [SENTENCE]	21 %	17 %	19 %
[QA] // [TEXTBLOCK]	23 %	12 %	20 %

TABLE 4 – % d'intersection entre les ensembles de similarités

différence, nous considérons l'ensemble de liens produits par une question comme une entité unique et calculons l'intersection avec l'ensemble de liens générés par d'autres méthodes de similarité.

Pour cela, nous avons réalisé le même appariement que celui présenté dans la section 4 (noté **QA** dans la suite), mais en utilisant la mesure de similarité uniquement sur le texte selon deux granularités : phrase (**SENTENCE**) et bloc de texte de l'OCR (**TEXTBLOCK**). Nous créons un lien entre deux phrases ou deux TextBlocks si la similarité entre leurs embeddings est inférieure à un seuil. Comme pour la similarité sur les questions, nous ne conservons que les N meilleurs liens. Dans cette expérience, N était fixé à 49.

Pour nous assurer que les liens sont comparés au même niveau de granularité, nous considérons que les liens sont identiques s'ils pointent vers la même page. Le pourcentage de chevauchement est calculé comme l'intersection entre l'ensemble des liens produits par les questions sur une phrase ou un TextBlock et ceux produits par la similarité sur le texte.
Bien que cette évaluation (Table 4) ne permette pas à elle seule de mesurer la qualité de nos liens, elle nous montre que notre méthode produit des liens originaux avec près de 80% des 49 pages les plus similaires qui sont différentes de celles produites en utilisant des méthodes de similarité directement sur les segments de texte. Une évaluation subjective qui vérifiera le retour des lecteurs professionnels aux liens et explications proposés par notre méthode sera bientôt réalisée.

8 Conclusion

Cet article propose une nouvelle approche pour construire des liens entre des documents en se basant sur la génération de questions. Nos expériences montrent la qualité de nos questions générées automatiquement, leur pertinence dans un contexte local, ainsi que l'originalité des liens produits par l'appariement de ces questions. Des expériences restent à mener pour étudier de manière plus qualitative les liens générés, ainsi que pour enrichir et filtrer de manière plus fine la grande quantité de questions sur le corpus.

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 2022-AD011012688R1 attribuée par GENCI.

Références

ALBERTI C., ANDOR D., PITLER E., DEVLIN J. & COLLINS M. (2019). Synthetic QA corpora generation with roundtrip consistency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 6168–6173, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1620.

BECHET F., ANTOINE E., AUGUSTE J. & DAMNATI G. (2022). Question generation and answering for exploring digital humanities collections. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 4561–4568, Marseille, France : European Language Resources Association.

BIRD S., KLEIN E. & LOPER E. (2009). *Natural language processing with Python : analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

DAI A. M., OLAH C. & LE Q. V. (2015). Document embedding with paragraph vectors. *CoRR*, **abs/1507.07998**.

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.

KAMAL EDDINE M., TIXIER A. & VAZIRGIANNIS M. (2021). BARThez : a skilled pretrained French sequence-to-sequence model. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 9369–9390, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.740.

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. DOI : 10.18653/v1/2020.acl-main.645.

PALMER M., GILDEA D. & KINGSBURY P. (2005). The proposition bank : An annotated corpus of semantic roles. *Comput. Linguist.*, **31**(1), 71–106. DOI : 10.1162/0891201053630264.

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.

REIMERS N. & GUREVYCH I. (2019). Sentence-bert : Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* : Association for Computational Linguistics.

REIMERS N. & GUREVYCH I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 4512–4525, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.365.

TSATSARONIS G., VARLAMIS I. & VAZIRGIANNIS M. (2014). Text relatedness based on a word thesaurus. *CoRR*, **abs/1401.5699**.

WANG Z., MI H. & ITTYCHERIAH A. (2016). Sentence similarity learning by lexical decomposition and composition. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics : Technical Papers*, p. 1340–1349, Osaka, Japan : The COLING 2016 Organizing Committee.

Annexes

A Exemples de questions par annotations en likert

Le score donné à la question pour l'affirmation est indiqué entre parenthèses.
"La question est pertinente dans le contexte général de la lecture" :

- Qu'est ce qui ne sautent pas aux yeux ? (1)
- Comment simplifient-t-on le problème ? (1)
- Qui a rejeté la solution préconisée ? (1)
- Qu'est ce qui n'existe pas ? (1)
- Qu'exige le jeu du marché ? (3)
- Qu'a permis ce système dans la période en question ? (3)
- Qui est à l'origine de la participation des travailleurs ? (3)
- Qu'exige l'autogestion ? (5)
- Qui agit en tant que force idéologique d'avant-garde dans la réalisation des intérêts fondamentaux des travailleurs ? (5)
- Contre quoi les travailleurs ont-ils à lutter ? (5)
- Quel est le moyen le plus simple d'exercer un contrôle sur les travailleurs ? (5)

"La question est pertinente dans le contexte de la phrase" :

- Quelle est la conséquence de ces déclarations ? (1)
- Qu'accompagne l'autogestion ? (1)
- Qu'est-ce qui est à l'origine de l'économie auto-organisée et de ses intérêts ? (1)
- Quel syndicat s'oppose à ce mouvement ? (1)
- Qui favorise le développement de ce mouvement ? (3)
- Qui tend à atomiser l'économie ? (3)
- Combien de travailleurs sont capables de gérer l'usine ? (3)
- Qu'exige le jeu du marché ? (5)
- Quelle proportion des travailleurs se retrouve salariée ? (5)
- Que permet le statut de coopérative ? (5)
- Quel est le moyen le plus simple d'exercer un contrôle sur les travailleurs ? (5)