# Un jeu de données pour répondre à des questions visuelles à propos d'entités nommées en utilisant des bases de connaissances

Paul Lerner, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, Jose Moreno, Jesús Lovón-Melgarejo

# Un jeu de données pour répondre à des questions visuelles à propos d'entités nommées en utilisant des bases de connaissances

Paul Lerner[1] Olivier Ferret[2] Camille Guinaudeau[1] Hervé Le Borgne[2]
Romaric Besançon[2] Jose G Moreno[3] Jesús Lovón Melgarejo[3]
(1) Université Paris-Saclay, CNRS, LISN, 91400, Orsay, France
(2) Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
(3) IRIT, Université Paul Sabatier, Toulouse, France
prenom.nom@lisn.upsaclay.fr, prenom.nom@cea.fr

RÉSUMÉ

Dans le contexte général des traitements multimodaux, nous nous intéressons à la tâche de réponse à des questions visuelles à propos d'entités nommées en utilisant des bases de connaissances (KVQAE). Nous mettons à disposition ViQuAE, un nouveau jeu de données de 3 700 questions associées à des images, annoté à l'aide d'une méthode semi-automatique. C'est le premier jeu de données de KVQAE comprenant des types d'entités variés associé à une base de connaissances composée d'1,5 million d'articles Wikipédia, incluant textes et images. Nous proposons également un modèle de référence de KVQAE en deux étapes : recherche d'information puis extraction des réponses. Les résultats de nos expériences démontrent empiriquement la difficulté de la tâche et ouvrent la voie à une meilleure représentation multimodale des entités nommées.

ABSTRACT

**ViQuAE, a Dataset for Knowledge-based Visual Question Answering about Named Entities**

In the context of multimodal processing, we focus our work on Knowledge-based Visual Question Answering about named Entities (KVQAE). We provide ViQuAE, a novel dataset of 3,700 questions paired with images, annotated using a semi-automatic method. It is the first KVQAE dataset to cover a wide range of entity types, associated with a knowledge base composed of 1.5M Wikipedia articles paired with images. To set a baseline on the benchmark, we address KVQAE as a two-stage problem : Information Retrieval and Extractive Reading Comprehension. The experiments empirically demonstrate the difficulty of the task and pave the way towards better multimodal entity representations.

MOTS-CLÉS : jeu de données, question-réponse visuelle, bases de connaissances, multimodalité.

KEYWORDS: dataset, knowledge-based visual question answering, multimodality.

## 1 Introduction et travaux connexes

La fusion de modalités telles que l'image et le texte pour rechercher des informations est un problème ancien et difficile du fait de la différence de niveaux de leurs sémantiques (Srihari *et al.*, 2000). C'est particulièrement vrai pour répondre à des questions visuelles à propos d'entités nommées en utilisant des bases de connaissances (KVQAE, *Knowledge-based Visual Question Answering about named*

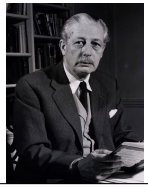
| Requête (entrée) | | Article pertinent dans la base de connaissances | |
|---|---|---|---|
|  | *"Which constituency did this man represent when he was Prime Minister ?"* |  | "Macmillan indeed lost Stockton in the landslide Labour victory of 1945, but returned to Parliament in the November 1945 by-election in **Bromley**." |
|  | "*In which year did this ocean liner make her maiden voyage ?*" |  | "Queen Elizabeth 2, often referred to simply as QE2, is a floating hotel and retired ocean liner built for the Cunard Line which was operated by Cunard as both a transatlantic liner and a cruise ship from **1969** to 2008." |

FIGURE 1 – Exemple de questions du jeu de données ViQuAE avec leur image contextuelle et la source de la réponse (issue de la base de connaissances).

*Entities*), la tâche considérée dans cet article, où différents types de relations peuvent lier une question et l'image qui lui est associée comme contexte (cf. Figure 1).

Dans la tâche classique de réponse à des questions visuelles (VQA, *Visual Question Answering*), le contenu de l'image associée, par exemple la couleur d'un objet ou le nombre d'objets, est le sujet de la question (Antol *et al.*, 2015). La VQA fondée sur les connaissances (Wang *et al.*, 2017, 2018; Marino *et al.*, 2019) utilise quant à elle l'image comme contexte pour poser des questions et trouver des réponses dans des bases de connaissances (BC). Cependant, ces deux champs de recherche se focalisent principalement sur des catégories d'objets à gros grain en s'appuyant sur un prétraitement de détection d'objets (Anderson *et al.*, 2018; Gardères & Ziaeefard, 2020). Dans cette optique, la seconde question de la Figure 1 pourrait porter sur le type de bateau : "*Est-ce un bateau de pêche ?*" Au contraire, notre travail se concentre sur des questions nécessitant des connaissances à propos des entités nommées, telles que le bateau *Queen Elizabeth 2*. Nous avons conçu et publions le jeu de données ViQuAE dans ce but [1]. Notre jeu de données a été conçu comme un benchmark pour suivre les progrès des systèmes de KVQAE. En effet, nous pensons que la KVQAE est une tâche bien définie qui peut être évaluée facilement. Elle est donc appropriée pour rendre compte des progrès de la qualité des représentations multimodales d'entités nommées. La représentation multimodale des entités est une question centrale qui permettra de rendre les interactions homme-machine plus naturelles. Par exemple, en regardant un film, on peut se demander "*Où ai-je déjà vu cette actrice ?*" ou "*Est-ce qu'elle a déjà gagné un Oscar ?*" Les questions sur les entités nommées sont très difficiles, car les BC actuelles en contiennent des millions. De ce point de vue, utiliser chaque modalité indépendamment n'est pas suffisamment discriminant pour répondre au besoin de l'utilisateur. À titre d'exemple, dans les images de la Figure 1, il est assez complexe de reconnaître *Harold Macmillan* au sein d'une BC contenant des millions de *personnes*. Cependant, on peut déduire de la question qu'il était *premier ministre*, ce qui permet de filtrer les candidats à quelques centaines.

Shah *et al.* (2019) ont déjà travaillé sur la KVQAE mais se sont limités aux entités nommées de type personne. Au contraire, ViQuAE comprend divers types d'entités. Cette diversité est une question centrale dans la KVQAE, notamment en raison de l'hétérogénéité des représentations visuelles qui en résulte. Entre autres entités, les entreprises peuvent être ainsi représentées par un bâtiment (*e.g.* leur siège), un produit manufacturé qu'elles vendent ou simplement leur logo. La KVQAE nécessite donc une représentation multimodale des connaissances, ce qui la distingue clairement de la recherche

1. Disponible via https://github.com/PaulLerner/ViQuAE.

d'image par le contenu. Cette diversité implique également la nécessité d'étudier d'autres types d'entités que les personnes, qui peuvent assez bien être reconnues visuellement à partir de leur seul visage. Par ailleurs, les questions du jeu de données de Shah *et al.* sont générées automatiquement à partir de patrons et de Wikidata, ce qui limite en particulier leur diversité de formes et de sujets. Enfin, Shah *et al.* utilisent une base de connaissances construite à partir d'un graphe de connaissances au lieu d'un texte non structuré, ce qui est aussi une différence importante avec notre travail.

Sur un autre plan, ViQuAE, avec ses 3 700 questions, s'inscrit dans le courant des travaux sur l'apprentissage sans (*zero-shot)* ou avec peu d'exemples (*few-shot*), avec une double idée : d'une part, la diversité des tâches unissant texte et image ne permet pas de développer des jeux de données d'une taille suffisante pour entraîner de gros modèles à partir de zéro ; d'autre part, les percées des travaux reposant sur les *Foundation Models* (Bommasani *et al.*, 2021) permettent de s'affranchir d'un tel entraînement. Nous espérons ainsi que ViQuAE encouragera les études vers des modèles transférables ou des techniques d'apprentissage sans ou avec peu d'exemples, nécessaires pour la KVQAE.

Plus spécifiquement, nous présentons dans cet article trois principales contributions : (i) nous fournissons un nouveau jeu de données pour la KVQAE, le premier à inclure divers types d'entités ainsi qu'une procédure extensible pour l'annotation semi-automatique ; (ii) nous rendons disponible une BC multimodale d'1,5 million d'entités fondée sur Wikipédia ; (iii) nous proposons et mettons en libre accès des méthodes d'apprentissage avec peu ou sans exemple pour traiter la KVQAE, étant les premiers à traiter la tâche sur divers types d'entités et en utilisant une BC textuelle.

# 2 Jeu de données et base de connaissances ViQuAE

## 2.1 Annotation automatique

Pour limiter les efforts d'annotation manuelle, nous nous appuyons sur des jeux de données de question-réponse (QA) existants, qui comprennent des questions couvrant divers sujets et entités. Nous avons ainsi décidé d'utiliser le jeu de données textuel TriviaQA en raison de sa taille et de la typologie de ses questions (Joshi *et al.*, 2017). L'idée principale du processus est de remplacer la mention de l'entité dans la question par une représentation visuelle de l'entité. Celle-ci est alors référencée par une mention ambiguë (*e.g. "cet homme"*). De cette façon, il n'est pas possible de répondre à la question sans s'appuyer sur l'image contextuelle. Dans le premier exemple de la Figure 1, la mention de l'entité nommée "*Harold Macmillan*" de la question originale est ainsi remplacée par la mention ambiguë "*this man*".

Ce processus débute par une analyse syntaxique et une identification des entités nommées dans les questions à l'aide de `spaCy` (Explosion, 2022). À partir de ces mentions d'entité, puisque la réponse à la question est connue, la désambiguïsation peut être effectuée en vérifiant si la réponse est présente dans l'article Wikipédia de l'entité candidate. Wikidata permet de recueillir des informations sur les entités désambiguïsées : leur type, leur profession, leur genre et leur catégorie Commons. Cette dernière est utilisée pour trouver une image pertinente tandis que les autres sont nécessaires pour générer une mention ambiguë. Les humains sont mentionnés par leur profession et les autres entités par leur type. De plus, si le genre est disponible, nous utilisons également "*this man/woman*" et "*he-him-his/she-her-hers*" selon la dépendance syntaxique de la mention originale. Étant donné que certaines entités abstraites, telles que les pays ou les nationalités, sont souvent mentionnées dans les questions mais ne sont pas pertinentes pour la KVQAE, le type d'entité est restreint à faire partie ou

être une sous-classe d'une liste de types construite manuellement, disponible avec le jeu de données. De plus, pour se conformer à la RGPD (European Parliament, 2016), seules les questions portant sur des personnes décédées sont conservées. Les images sont récupérées à partir de la catégorie Commons de l'entité.

## 2.2 Annotation manuelle

L'annotation automatique décrite ci-dessus présente quelques inconvénients. Les deux principales sources d'erreurs sont : (i) l'image sélectionnée, qui peut être inappropriée ; (ii) la spécificité de la question, qui permet parfois de répondre sans regarder l'image. Pour remédier à ce problème, une interface d'annotation a été conçue à l'aide de Label Studio (Tkachenko *et al.*, 2021). L'annotateur peut reformuler librement la question tant que la réponse n'est pas modifiée. Il doit également choisir parmi huit images candidates si celle sélectionnée n'est pas appropriée. En dernier recours, l'annotateur peut simplement rejeter la question. L'interface et les instructions d'annotation font partie de notre base de code.

Cette annotation manuelle a été réalisée par sept annotateurs internes. L'interface a permis de traiter environ 120 questions par heure. La proportion de questions à propos d'humains a été équilibrée pour assurer la diversité du jeu de données. Nous avons annoté 5 700 questions générées, parmi lesquelles 2 000 ont été écartées, principalement parce qu'elles étaient sur-spécifiées ou que l'image n'était pas pertinente. Par conséquent, le jeu de données ViQuAE est constitué de 3 700 questions, réparties aléatoirement en ensembles de taille égale pour l'entraînement, la validation et le test, sans recouvrement entre les images. La majorité (55 %) des questions valides ont été éditées par les annotateurs, avec une distance de Levenshtein moyenne de 5 mots.

Pour mesurer l'accord inter-annotateur, un sous-ensemble de 103 questions a été annoté par au moins 3 annotateurs différents. L'accord est ensuite calculé en utilisant le Kappa de Fleiss (Fleiss, 1971). Les annotateurs se sont mis d'accord pour rejeter ou non la question avec $\kappa = 0.33$, montrant un accord léger. En effet, déterminer si une question est sur-spécifiée ou non peut être assez subjectif. De plus, la reformulation de certaines questions sur-spécifiées peut être subtile. Cependant, il faut rappeler que, dans notre cas, le désaccord inter-annotateurs ne concerne pas la *réponse* à la question mais seulement le filtrage du jeu de données généré automatiquement, puisque les questions et les réponses sont définies dans TriviaQA et que l'annotateur *ne peut pas changer la réponse*.

## 2.3 Analyse des données

Le jeu de données ViQuAE se compose de 3 700 questions contextualisées par 3 300 images uniques, dont deux exemples sont présentés à la Figure 1. Les questions comportent en moyenne 12 mots, pour un vocabulaire de 4 700 mots. Sur les 3 700 réponses, les plus communes n'apparaissent que 13 fois, soit 0,3 % du total, ce qui montre l'absence de biais a priori sur les réponses. De plus, il n'y a qu'un chevauchement de 25 % des réponses et de 18 % des entités entre les ensembles d'entraînement et de test. Ces trois points soulignent la différence entre la KVQAE et la VQA (fondé sur la connaissance ou pas) et démontrent que traiter la KVQAE comme une tâche de classification serait inefficace.

Une contribution importante du jeu de données est sa diversité d'entités, qui est l'un des principaux défis pour les représentations multimodales (cf. Section 1). ViQuAE comprend près de mille types d'entités différents (issus de l'ontologie Wikidata) parmi ses 2 400 entités uniques. Toutefois, ces types

| | ViQuAE | KVQA |
|---|---|---|
| # Questions | 3 700 | 183K |
| # Questions par image | 1,1 | 7,4 |
| Vocabulaire | 4,7K | 0,6K* |
| Longueur moyenne des questions | 12,4 | 10,1 |
| Réponse la plus probable a priori | 0,3 % | 15,9 % |
| Chevauchement entre les réponses | 25,3 % | 89,4 % |
| Chevauchement entre les entités | 18,1 % | 40,6 % |
| # Questions par entité | 1,5 | 9,7 |
| # Types d'entités | 980 | 1 |

TABLE 1 – Statistiques du jeu de données par rapport à KVQA (Shah *et al.*, 2019). *En moyenne sur 49 sous-ensembles aléatoires de la même taille que ViQuAE, le vocabulaire du jeu de données KVQA entier se compose de 8,4K tokens.

ne sont pas exclusifs : 1,6 type sont attribués à chaque entité en moyenne. Le jeu de données comporte 43 % d'humains, sans prendre en compte d'autres entités semblables, comme les personnages fictifs ou mythologiques, ou des groupes d'humains, par exemple les groupes de musique. Un résumé des statistiques comparées avec le jeu de données KVQA de Shah *et al.* (2019) est reporté dans le Tableau 1. Nous pouvons constater que, malgré sa petite taille, ViQuAE est plus diversifié sous certains aspects. Cependant, le jeu de données ViQuAE présente aussi certaines limites. L'un des inconvénients de notre processus d'annotation, et plus précisément de la désambiguïsation des entités nommées, est que les réponses sont garanties de se trouver dans la page Wikipédia de l'entité, *i.e.*, les questions sont *single-hop* au niveau du document. Bien sûr, la question peut toujours nécessiter un raisonnement sur plusieurs phrases ou paragraphes du document. En revanche, (Shah *et al.*, 2019) comprend plusieurs questions *multi-hop* qui, même si elles ne semblent pas très naturelles, permettent d'évaluer les capacités de raisonnement du modèle.

### 2.4 La base de connaissances ViQuAE

La BC est construite à partir de la sauvegarde du 01/08/2019 de Wikipédia, disponible dans KILT (Petroni *et al.*, 2021), comprenant 5,9M d'articles. Chacun d'eux est associé à une entité Wikidata. Pour obtenir une représentation visuelle de l'entité, une image unique est extraite de Wikidata, dans l'ordre suivant de préférence des propriétés Wikidata : (i) P18 "image" ; (ii) P154 "image du logotype" ; (iii) P41 "image du drapeau" ; (iv) P94 "image du blason" ; (v) P2425 "ruban de médaille". Les articles sans image sont écartés, ce qui aboutit à une BC d'1,5 million d'articles, dont 542 000 à propos d'humains, chacun associé à une image. C'est supérieur à l'échelle des expériences de Shah *et al.* (2019) de plus de deux ordres de grandeur. 95 % des images de la base de connaissances sont uniques.

## 3 Expériences

Nous traitons le problème de la KVQAE en deux étapes : recherche d'information (RI) puis extraction des réponses (*extractive reading comprehension*). Cette décomposition en deux étapes est standard en

QA textuelle (e.g. Chen *et al.*, 2017). D'après Joshi *et al.* (2017), les alias Wikipédia d'une réponse donnée sont considérés comme des réponses valides.

## 3.1 Recherche d'information

**Recherche de texte** Nous adoptons une approche de fusion tardive : la recherche est effectuée indépendamment avec la question et l'image puis les résultats sont fusionnés au niveau des scores. En amont de la recherche, nous enlevons les données semi-structurées des articles, comme les tableaux et les listes. Chaque article est ensuite divisé en passages disjoints de 100 mots tout en préservant les limites des phrases, ce qui produit 12 millions de passages. Le titre de l'article est concaténé au début de chaque passage. Nous utilisons BM25 (Robertson *et al.*, 1995) et DPR (Karpukhin *et al.*, 2020) pour définir une référence *zero-shot* et *few-shot*, respectivement. DPR est d'abord pré-entraîné sur TriviaQA, filtré de toutes les questions utilisées dans ViQuAE, avant d'être ajusté sur ViQuAE. Nous considérons également le modèle sans ajustement, entraîné uniquement sur TriviaQA, comme une autre référence *zero-shot*.

**Recherche d'image** Pour la recherche d'images, nous utilisons deux représentations différentes de manière exclusive : ArcFace (Deng *et al.*, 2019) pour les visages, si au moins un visage est détecté ; ImageNet-ResNet (He *et al.*, 2016) et CLIP (Radford *et al.*, 2021) pour l'image complète. Par conséquent, la BC est divisée en deux parties : les humains avec un visage détecté et les non-humains en considérant que les visages ne sont pertinents que pour les entités humaines. Comme Deng *et al.* (2019), nous utilisons MTCNN (Zhang *et al.*, 2016) pour la détection des visages. Si plusieurs visages sont détectés, seul celui associé à la plus forte probabilité est conservé. 6,6 % des humains de la BC n'ont pas de visage détecté et ont donc été écartés. ArcFace est pré-entraîné sur MS-Celeb (Guo *et al.*, 2016), composé de photos de célébrités. Ses entités ont un certain chevauchement avec ViQuAE, qui est analysé dans la section suivante.

**Fusion multimodale** Les résultats de la recherche par l'image sont ensuite mis en correspondance avec les passages pour la fusion avec la recherche textuelle. Les scores des résultats de ces modèles ayant des distributions très différentes, ils sont centrés-réduits avant de les fusionner. La fusion est faite via une combinaison linéaire : $P = \alpha_b B + \alpha_d D + \mathbf{F}\alpha_a A + (1-\mathbf{F})(\alpha_i I + \alpha_c C)$. On note $B, D, A, I, C$ les scores BM25, DPR, ArcFace, ImageNet-ResNet et CLIP respectivement, chacun étant pondéré par l'hyperparamètre $\alpha_j$. $\mathbf{F} \in \{0,1\}$ dénote la détection d'un visage. Seuls les 100 premiers passages sont considérés. Par conséquent, si, compte tenu d'une requête, un passage n'est pas retrouvé par un système donné, on lui attribue le score minimum des autres passages retrouvés par ce système. Les passages sont ensuite réordonnés par rapport au score $P$. Les hyperparamètres d'interpolation $\alpha_j$ sont réglés sur l'ensemble de validation en utilisant une recherche par quadrillage pour maximiser le rang réciproque moyen. Pour limiter l'espace de recherche et permettre une comparaison directe entre BM25 et DPR nous contraignons $\sum_j \alpha_j = 1$ et n'utilisons qu'un seul modèle pour la recherche texte : on a donc $\alpha_b = 0$ ou $\alpha_d = 0$.

**Résultats** Puisqu'il est fondé sur TriviaQA (Joshi *et al.*, 2017), ViQuAE n'est supervisé que de façon distante, *i.e.* un passage est jugé pertinent s'il contient la réponse. Nous évaluons la RI avec la précision à K (P@K) et le rang réciproque moyen (MRR) ainsi que Hits@K. Hits@K représente la

| # | Modèle | MRR | P@1 | P@20 | Hits@20 |
|---|---|---|---|---|---|
| a | $B$ (BM25, texte seulement) | 19,0 | 13,1 | 5,9 | 39,5 |
| b | $D_0$ (DPR *zero-shot*, texte seulement) | $30{,}5^{a}$ | $21{,}2^{a}$ | $16{,}2^{ac}$ | $60{,}5^{ac}$ |
| c | $0{,}3(B+\mathbf{F}A)+(1-\mathbf{F})(0{,}1I+0{,}3C)$ | $27{,}9^{a}$ | $20{,}4^{a}$ | $10{,}1^{a}$ | $50{,}5^{a}$ |
| d | $0{,}3(D_0+\mathbf{F}A)+(1-\mathbf{F})(0{,}1I+0{,}3C)$ | $36{,}0^{abce}$ | $26{,}7^{abce}$ | $17{,}1^{ac}$ | $65{,}2^{abce}$ |
| e | $D_f$ (DPR *few-shot*, texte seulement) | $32{,}8^{abc}$ | $22{,}8^{a}$ | $16{,}4^{ac}$ | $61{,}2^{ac}$ |
| f | $0{,}3(D_f+\mathbf{F}A)+0{,}2(1-\mathbf{F})(I+C)$ | $\mathbf{37{,}9}^{abcde}$ | $\mathbf{27{,}8}^{abce}$ | $\mathbf{17{,}5}^{ac}$ | $\mathbf{65{,}7}^{abce}$ |

TABLE 2 – Résultats de la RI avec les baselines textuelles et la fusion de la recherche multimodale, dans les deux configurations d'apprentissage : sans ou avec peu d'exemples. Les exposants dénotent des différences significatives dans le test de randomisation de Fisher (Fisher, 1937; Smucker *et al.*, 2007) avec $p \leq 0{,}01$. Hits@1 est omis car il est équivalent à P@1.

proportion de questions pour lesquelles la RI récupère *au moins un* passage pertinent parmi les K premiers. Les résultats sont présentés dans le Tableau 2. Nous présentons également comme références les performances de BM25 et de DPR utilisant seulement le texte. Le gain de performance de DPR par rapport à BM25 est important, même dans sa version *zero-shot* où il surpasse significativement BM25 et même la recherche multimodale fondée sur BM25 en P@20 et Hits@20. Contrairement à BM25, DPR est capable de trouver des passages pertinents même avec très peu de chevauchement lexical grâce à ses représentations sémantiques abstraites. Néanmoins, il faut noter que la fusion multimodale apporte également des gains de performance significatifs. Ce gain diffère selon le type de l'entité-sujet de la question. Pour les questions à propos d'humains, la P@1 passe de 14,4 avec BM25 seul à 24,4 en fusionnant BM25 et la recherche d'images, soit une amélioration de 70 %. En comparaison, l'amélioration de 41 % de la P@1 pour les questions sur les non-humains est plus faible. En outre, sur le sous-ensemble d'entités qui se chevauchent avec MS-Celeb (le jeu de données de pré-entraînement d'ArcFace), P@1 augmente encore à 25,7, ce qui représente une amélioration de 5 % par rapport à tous les humains. La tendance est similaire avec DPR, bien que sa baseline textuelle soit meilleure.

## 3.2 Extraction des réponses

**Méthodes** Pour établir notre référence sur ViQuAE, nous nous limitons à un modèle textuel car nous faisons l'hypothèse qu'une fois le passage pertinent retrouvé en associant texte et image, il est possible de répondre à la question sans utiliser l'image (cf. par exemple la Figure 1). L'extraction des réponses est réalisée avec le modèle BERT multi-passage de Wang *et al.* (2019). Nous laissons le ré-ordonnancement multimodal pour de futurs travaux mais nous avons expérimenté la pondération du score de réponse $a$ avec le score de RI du passage $P$ t.q. $a \leftarrow a \cdot P$ (Wang *et al.*, 2019). Les mêmes hyperparamètres que Karpukhin *et al.* (2020) sont utilisés, à l'exception du ratio de passages pertinents et non pertinents par question, qui est fixé à $8:16$. À l'inférence, l'extraction est appliquée sur les 24 premiers résultats de la RI. Comme à la section précédente, le modèle est d'abord pré-entraîné sur notre sous-ensemble de TriviaQA, puis ajusté sur ViQuAE.

**Résultats** Les résultats sont présentés dans le Tableau 3. Ils sont globalement assez faibles par rapport à l'état de l'art en Question-Réponse textuelle. Pour mieux comprendre ces chiffres, nous avons étudié deux configurations différentes : (i) *mi-oracle*, où les 24 premiers résultats de la RI sont

| # Exemples | Configuration | F1 | Appariement exact |
|---|---|---|---|
| Aucun | Texte seulement | 20,96 | 18,06 |
| Aucun | + pondération RI | 21,19 | 18,22 |
| Peu | Texte seulement | $25{,}43 \pm 0{,}42$ | $22{,}07 \pm 0{,}54$ |
| Peu | + pondération RI | $25{,}50 \pm 0{,}38$ | $22{,}10 \pm 0{,}54$ |
| Peu | Mi-oracle | $44{,}10 \pm 0{,}39$ | $40{,}32 \pm 0{,}43$ |
| Peu | Oracle complet | $63{,}17 \pm 1{,}18$ | $57{,}55 \pm 1{,}10$ |

TABLE 3 – Résultats de l'extraction de réponses sur l'ensemble de test de ViQuAE, moyenné après entraînement avec 5 graines aléatoires différentes pour le modèle *few-shot*. Les modèles *zero* et *few-shot* partagent les mêmes résultats de RI à l'inférence (24 premiers passages).

filtrés pour ne contenir que des passages pertinents ; (ii) *oracle complet*, où le modèle ne reçoit que des passages pertinents. Ces résultats oracles pourraient servir de référence haute aux futures études.

# 4 Conclusion et perspectives

Nous présentons un nouveau jeu de données, ViQuAE, conçu comme un cadre d'évaluation pour suivre les progrès des systèmes de KVQAE. ViQuAE a été annoté selon une procédure semi-automatique que nous fournissons également. Ses questions ont pour cible une base de connaissances librement disponible d'1,5 million d'articles Wikipédia associés à des images. Nous proposons une approche de la KVQAE en deux étapes, distinguant recherche d'information et extraction des réponses, avec des méthodes d'apprentissage sans ou avec peu d'exemples dans les deux cas. Un résultat notable de cette première référence est l'apport positif de l'association du texte et de l'image dans ces différentes configurations. Sans négliger l'extraction des réponses, les évaluations soulignent par ailleurs la nécessité d'une meilleure RI. En effet, notre stratégie de fusion tardive néglige l'interaction entre les modalités. Les travaux futurs devront se concentrer sur une meilleure représentation multimodale, idéalement en intégrant le texte et l'image dans le même espace, tant du côté de la requête que du côté de la BC. Une attention particulière devra être accordée à la représentation des entités non-humaines. Ces représentations multimodales pourront aussi bénéficier à l'étape d'extraction des réponses car nos expériences montrent que l'utilisation d'un modèle textuel seul est insuffisante si la RI est bruitée. D'autre part, bien que nous ayons démontré l'efficacité de notre BC, on pourrait bénéficier d'une BC plus riche visuellement, avec plusieurs images par entité, afin de prendre en compte la diversité des représentations. Nous espérons que ce travail encouragera la recherche vers une meilleure représentation multimodale des entités nommées.

# Remerciements

Nous remercions les relecteurs de TALN pour leurs précieuses suggestions. Ce travail a été financé par le projet ANR-19-CE23-0028 MEERQAT. Ce travail a bénéficié d'un accès aux moyens de calcul de l'IDRIS au travers de l'allocation de ressources 2021-AD011012846 attribuée par GENCI.

# Références


ANDERSON P., HE X., BUEHLER C., TENEY D., JOHNSON M., GOULD S. & ZHANG L. (2018). Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

ANTOL S., AGRAWAL A., LU J., MITCHELL M., BATRA D., ZITNICK C. L. & PARIKH D. (2015). VQA : Visual Question Answering. In *2015 IEEE International Conference on Computer Vision (ICCV)*, p. 2425–2433, Santiago, Chile : IEEE. DOI : 10.1109/ICCV.2015.279.

BOMMASANI R., HUDSON D. A., ADELI E., ALTMAN R., ARORA S., VON ARX S., BERNSTEIN M. S., BOHG J., BOSSELUT A., BRUNSKILL E., BRYNJOLFSSON E., BUCH S., CARD D., CASTELLON R., CHATTERJI N., CHEN A., CREEL K., DAVIS J. Q., DEMSZKY D., DONAHUE C., DOUMBOUYA M., DURMUS E., ERMON S., ETCHEMENDY J., ETHAYARAJH K., FEI-FEI L., FINN C., GALE T., GILLESPIE L., GOEL K., GOODMAN N., GROSSMAN S., GUHA N., HASHIMOTO T., HENDERSON P., HEWITT J., HO D. E., HONG J., HSU K., HUANG J., ICARD T., JAIN S., JURAFSKY D., KALLURI P., KARAMCHETI S., KEELING G., KHANI F., KHATTAB O., KOH P. W., KRASS M., KRISHNA R., KUDITIPUDI R., KUMAR A., LADHAK F., LEE M., LEE T., LESKOVEC J., LEVENT I., LI X. L., LI X., MA T., MALIK A., MANNING C. D., MIRCHANDANI S., MITCHELL E., MUNYIKWA Z., NAIR S., NARAYAN A., NARAYANAN D., NEWMAN B., NIE A., NIEBLES J. C., NILFOROSHAN H., NYARKO J., OGUT G., ORR L., PAPADIMITRIOU I., PARK J. S., PIECH C., PORTELANCE E., POTTS C., RAGHUNATHAN A., REICH R., REN H., RONG F., ROOHANI Y., RUIZ C., RYAN J., RÉ C., SADIGH D., SAGAWA S., SANTHANAM K., SHIH A., SRINIVASAN K., TAMKIN A., TAORI R., THOMAS A. W., TRAMÈR F., WANG R. E., WANG W., WU B., WU J., WU Y., XIE S. M., YASUNAGA M., YOU J., ZAHARIA M., ZHANG M., ZHANG T., ZHANG X., ZHANG Y., ZHENG L., ZHOU K. & LIANG P. (2021). On the Opportunities and Risks of Foundation Models. *arXiv :2108.07258 [cs]*. arXiv : 2108.07258.

CHEN D., FISCH A., WESTON J. & BORDES A. (2017). Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1870–1879.

DENG J., GUO J., XUE N. & ZAFEIRIOU S. (2019). ArcFace : Additive Angular Margin Loss for Deep Face Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

EUROPEAN PARLIAMENT T. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). Legislative Body : EP, CONSIL.

EXPLOSION (2022). Spacy : Industrial-strength NLP. https://spacy.io/.

FISHER R. A. (1937). *The Design of Experiments*. Oliver & Boyd, Edinburgh & London., 2ème édition.

FLEISS J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, **76**(5), 378–382. DOI : 10.1037/h0031619.

GARDÈRES F. & ZIAEEFARD M. (2020). ConceptBert : Concept-Aware Representation for Visual Question Answering. *Findings of the Association for Computational Linguistics : EMNLP 2020*, p. 489–498.

GUO Y., ZHANG L., HU Y., HE X. & GAO J. (2016). MS-Celeb-1M : A Dataset and Benchmark for Large-Scale Face Recognition. In B. LEIBE, J. MATAS, N. SEBE & M. WELLING, Éds., *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, p. 87–102, Cham : Springer International Publishing. DOI : 10.1007/978-3-319-46487-9_6.

HE K., ZHANG X., REN S. & SUN J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 770–778.

JOSHI M., CHOI E., WELD D. & ZETTLEMOYER L. (2017). TriviaQA : A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1601–1611, Vancouver, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/P17-1147.

KARPUKHIN V., OGUZ B., MIN S., LEWIS P., WU L., EDUNOV S., CHEN D. & YIH W.-T. (2020). Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 6769–6781, Online : Association for Computational Linguistics.

MARINO K., RASTEGARI M., FARHADI A. & MOTTAGHI R. (2019). OK-VQA : A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, p. 3195–3204.

PETRONI F., PIKTUS A., FAN A., LEWIS P., YAZDANI M., DE CAO N., THORNE J., JERNITE Y., KARPUKHIN V., MAILLARD J., PLACHOURAS V., ROCKTÄSCHEL T. & RIEDEL S. (2021). KILT : a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 2523–2544, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.naacl-main.200.

RADFORD A., KIM J. W., HALLACY C., RAMESH A., GOH G., AGARWAL S., SASTRY G., ASKELL A., MISHKIN P., CLARK J. *et al.* (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, p. 8748–8763 : PMLR.

ROBERTSON S. E., WALKER S., JONES S., HANCOCK-BEAULIEU M. M. & GATFORD M. (1995). Okapi at TREC-3. In D. K. HARMAN, Éd., *Third Text REtrieval Conference (TREC-3)*, volume 500-225 de *NIST Special Publication*, p. 109–126 : National Institute of Standards and Technology (NIST).

SHAH S., MISHRA A., YADATI N. & TALUKDAR P. P. (2019). KVQA : Knowledge-Aware Visual Question Answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, p. 8876–8884.

SMUCKER M. D., ALLAN J. & CARTERETTE B. (2007). A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, p. 623–632, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/1321440.1321528.

SRIHARI R. K., ZHANG Z. & RAO A. (2000). Intelligent Indexing and Semantic Retrieval of Multimodal Documents. *Information Retrieval*, **2**(2), 245–275. DOI : 10.1023/A :1009962928226.

TKACHENKO M., MALYUK M., SHEVCHENKO N., HOLMANYUK A. & LIUBIMOV N. (2021). Label Studio : Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

WANG P., WU Q., SHEN C., DICK A. & VAN DEN HENGE A. (2017). Explicit knowledge-based reasoning for visual question answering. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, p. 1290–1296.

WANG P., WU Q., SHEN C., DICK A. & VAN DEN HENGEL A. (2018). FVQA : Fact-Based Visual Question Answering. *IEEE transactions on pattern analysis and machine intelligence*, **40**(10), 2413–2427.

WANG Z., NG P., MA X., NALLAPATI R. & XIANG B. (2019). Multi-passage BERT : A Globally Normalized BERT Model for Open-domain Question Answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 5878–5882, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1599.

ZHANG K., ZHANG Z., LI Z. & QIAO Y. (2016). Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters*, **23**(10), 1499–1503. Conference Name : IEEE Signal Processing Letters, DOI : 10.1109/LSP.2016.2603342.