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

Augmentation de jeux de données de RI pour la recherche conversationnelle à initiative mixte

Pierre Erbacher[1] Philippe Preux[2] Jian-Yun Nie [3] Laure Soulier [1]
(1) Sorbonne University, ISIR, Paris
(2) INRIA Scool Team, Lille
(3) Montreal University, Canada
pierre.erbacher@isir.upmc.fr

RÉSUMÉ

Une des particularités des systèmes de recherche conversationnelle est qu'ils impliquent des initiatives mixtes telles que des questions de clarification générées par le système. L'évaluation de ces systèmes à grande échelle sur la tâche finale de RI est très difficile et nécessite des jeux de données adéquats contenant de telles interactions. Cependant, les jeux de données actuels se concentrent uniquement sur les tâches traditionnelles de RI ad hoc ou sur les tâches de clarification de la requête. Pour combler cette lacune, nous proposons une méthodologie pour construire automatiquement des jeux de données de RI conversationnelle à grande échelle à partir de jeux de données de RI ad hoc afin de faciliter les explorations sur la RI conversationnelle. Nous effectuons une évaluation approfondie montrant la qualité et la pertinence des interactions générées pour chaque requête initiale. Cet article montre la faisabilité et l'utilité de l'augmentation des jeux de données de RI ad-hoc pour la RI conversationnelle.

ABSTRACT

Augmenting Ad-Hoc IR Dataset for Mixed-initiative Conversational Search

A peculiarity of conversational search systems is that they involve mixed-initiatives such as system-generated query clarifying questions. Evaluating those systems at a large scale on the end task of IR is very challenging, requiring adequate datasets containing such interactions. However, current datasets only focus on either traditional ad-hoc IR tasks or query clarification tasks We propose a methodology to automatically build large-scale conversational IR datasets from ad-hoc IR datasets in order to facilitate explorations on conversational IR. We perform a thorough evaluation showing the quality and the relevance of the generated interactions for each initial query. This paper shows the feasibility and utility of augmenting ad-hoc IR datasets for conversational IR.

MOTS-CLÉS : Recherche conversationnelle, recherche d'informations, interactions mixtes, méthodologie de construction d'ensembles de données..

KEYWORDS: Conversational search, information retrieval, mixed-initiative interactions, dataset building methodology.

1 Introduction

Les systèmes conversationnels, y compris les assistants personnels et les chatbots, sont de plus en plus populaires pour une grande variété de tâches, notamment la recherche d'informations (RI) en ligne. Bien que des modèles de langue (ML) récents, comme ChatGPT d'OpenAI (Ouyang *et al.*, 2022), aient démontré leur capacité à répondre à des questions factuelles, ceux-ci ne peuvent pas

être considérés comme des systèmes de recherche conversationnelle. En effet, ces derniers sont entraînes à générer le texte le plus probable sans faire explicitement référence aux sources et sans garantie de véracité, ce qui amplifie les biais potentiels et les fausses vérités observés dans les données d'apprentissage (Bender *et al.*, 2021). Pour surmonter cette limitation, les systèmes de recherche conversationnelle doivent s'appuyer sur des capacités de recherche d'informations pour localiser les sources/documents pertinents (Shah & Bender, 2022; Dalton *et al.*, 2022; Zamani *et al.*, 2022; Anand *et al.*, 2020; Bender *et al.*, 2021). Pour améliorer la qualité des réponses, des modèles tels que LaMDA (Thoppilan *et al.*, 2022), WebGPT (Glaese *et al.*, 2022) ou Sparrow (Shuster *et al.*, 2022) peuvent additionnellement conditionner la génération de réponses à des informations récupérées par un outil de recherche d'information indépendant. Ceci ne garantit pas que les informations utilisées soit pertinentes, conduisant potentiellement à des réponses non véridiques ou non informatives (Nakano *et al.*, 2021). Comme le suggère (Dalton *et al.*, 2020a) il est important d'inclure les capacités de recherche d'information dans l'évaluation des modèles de recherche conversationnelle dans leur ensemble. Au-delà de la génération de réponses en langage naturel, l'une des principales capacités des systèmes de recherche conversationnelle est leur participation (pro)active à la conversation avec les utilisateurs afin de les aider à clarifier ou à affiner leurs besoins en information (initiative mixte). (Shah & Bender, 2022; Chu-Carroll & Brown, 1997; Dalton *et al.*, 2022; Zamani *et al.*, 2022; Anand *et al.*, 2020; Bender *et al.*, 2021; Radlinski & Craswell, 2017; Trippas *et al.*, 2020; Aliannejadi *et al.*, 2019; Keyvan & Huang, 2022; Zamani *et al.*, 2020b).

Les travaux pour évaluer la RI conversationnelle (Sekulić *et al.*, 2021; Aliannejadi *et al.*, 2019; Salle *et al.*, 2021; Bi *et al.*, 2021) se concentrent principalement sur l'évaluation de la qualité de la génération de questions de clarification à l'aide de jeux de données alignés, tels que Qulac (Aliannejadi *et al.*, 2019) et ClariQ (Aliannejadi *et al.*, 2021) qui contiennent des paires de requêtes et de questions de clarification, mais seulement sur 237 sujets différents. D'autres taches comme TREC CAsT (Dalton *et al.*, 2020b) se concentrent sur l'évaluation des capacités à retrouver les documents pertinents sans prendre en compte les interactions à initiative mixte. Ces limites démontrent la nécessité de construire des ensembles de données de RI à grande échelle contenant non seulement les requêtes de l'utilisateur mais aussi les interactions mixtes, utilisateur-système, mais aussi des signaux de pertinence. La collecte de telles données conversationnelles est un défi, en raison du coût élevé d'annotation.

Dans le contexte des systèmes de recommandation (Li *et al.*, 2018b; Zhou *et al.*, 2020a; Moon *et al.*, 2019; Kang *et al.*, 2019a; Liu *et al.*, 2021; Jia *et al.*, 2022) ont construit des jeux de données conversationnels en utilisant des annotateurs jouant le rôle de l'utilisateur et du système de recommandation autour d'objectifs et d'éléments prédéfinis. Grâce à la génération automatique, d'autres méthodes proposent de simuler les dialogues à partir de données de recommandation existantes (Gao *et al.*, 2022; Kang *et al.*, 2019b; Wu *et al.*, 2020; Zhou *et al.*, 2020b; Fu *et al.*, 2020). Les tentatives ci-dessus ont été possibles car les données de recommandation contiennent des articles annotés avec des catégories limitées et des caractéristiques discrètes telles que la couleur, la marque ou la gamme de prix. En RI, le transfert de ces approches est difficile dans la mesure où ces caractéristiques ne sont pas discrètes ou pas facilement identifiables, et les besoins en information sont beaucoup plus diversifiés. Les jeux de données disponibles sont très limités. Le jeu de données qulac fournit 10000 interactions à tour unique avec un tuple (intention, requête, question de clarification, réponse) sur 200 sujets seulement (Aliannejadi *et al.*, 2019). C'est loin d'être suffisant pour entraîner ou évaluer les techniques de RI conversationnelle dans divers contextes.

Dans notre travail, nous visons à générer automatiquement des interactions à initiatives mixtes entre un utilisateur et un système et proposer une méthodologie pour augmenter les jeux de données de RI ad-hoc. Pour ce faire, nous concevons un générateur de question de clarificationainsi qu'une

simulation utilisateur. Nous les utilisons pour générer des interactions à initiative mixte sur le jeu de données de RI MsMarco.

2 Etat de l'art

2.1 Évaluation de la recherche conversationnelle

L'évaluation des systèmes de RI conversationnelle reste un défi pour la communauté car cela implique d'évaluer la capacité du système à aider et guider l'utilisateur dans sa recherche. (Dalton *et al.*, 2022). Un tel système de recherche doit donc être capable de 1) générer des questions pour clarifier/expliquer les besoins en information des utilisateurs, et 2) récupérer des documents fournissant des informations pertinentes.

En ce qui concerne la tâche de question-réponse (QA), des jeux de données conversationnels (e.g., coQA (Reddy *et al.*, 2019)) ont été construits à partir de données QA "one-shot" telle que Squad (Rajpurkar *et al.*, 2018), Quac (Choi *et al.*, 2018), ELI5 (Fan *et al.*, 2019), ou OpenQA (Chen *et al.*, 2017) en évaluant aussi la capacité à retrouver les contextes pertinents pour répondre aux questions .

Malgré cette démarche intéressante, ces jeux de données sont insuffisants pour la RI car ils se concentrent généralement sur des questions factuelles au lieu de questions complexes ou exploratoires qui caractérisent les besoins en information.Le jeu de données TREC CAsT (Dalton *et al.*, 2020b) étend la portée des questions et aborde différentes facettes d'information au sein de la conversation (une facette peut être considérée comme une sous-catégorie spécifique du sujet). Cependant, le nombre de dialogues disponibles est très limité. D'autres jeux de données, tels que CANARD (Elgohary *et al.*, 2019), se concentrent sur le raffinement ou la reformulation des requêtes, sans interactions proactives de la part du système.

Le jeu de données MIMICS (Zamani *et al.*, 2020a) contient des questions de clarification de domaine ouvert à grande échelle recueillies auprès d'utilisateurs réels sur le moteur de recherche Bing.Cependant, ce jeu de données ne fournit pas de jugements de pertinence des documents ni d'interactions conversationnelles entre l'utilisateur et le système. Les jeux de données Qulac et ClariQ contiennent à la fois des jugements de pertinence de documents et les conversations mixtes associées. Ils sont construits à partir de la collection TREC Web Track 2009-12, qui fournit des paires de sujets et de facettes annotés, associés à des documents pertinents. Les réponses des utilisateurs ont été collectées par le biais de plateformes de crowd-sourcing. Cependant, la collecte de ces interactions a été coûteuse et les jeux de données restent petits avec seulement 237 sujets et 762 facettes de sujets. Cela est limité pour l'entraînement et l'évaluation des systèmes de recherche conversationnelle.

Face au manque de jeux de données adéquats, une idée de plus en plus répandue dans la communauté consiste à s'appuyer sur la simulation d'utilisateurs pour évaluer les systèmes de recherche conversationnelle : (Erbacher *et al.*, 2022; Salle *et al.*, 2021). Les simulations d'utilisateurs, qui consistent à imiter les requêtes et les réponses des utilisateurs, permettent d'évaluer diverses stratégies sans avoir de trajectoires de conversations prédéfinies dans les données. Par exemple, Salle et al (Salle *et al.*, 2021) évaluent leurs systèmes de clarification de requêtes avec une simulation utilisateur visant à générer des réponses. La simulation d'utilisateur est également exploitée pour concevoir des cadres d'évaluation pour les systèmes de recommandation conversationnels (Kang *et al.*, 2019b; Gao *et al.*, 2022; Wu *et al.*, 2020; Zhou *et al.*, 2020b; Fu *et al.*, 2020), donnant lieu à de grandes interactions de dialogue synthétique à partir de jeux de données de recommandation ad hoc. Cependant, dans le contexte de la recommandation, les conversations sont générées grâce a des contraintes de recherche explicites sur des caractéristiques annotées comme la gamme de prix, la couleur, l'emplacement, le

genre de film ou la marque (Asri *et al.*, 2016; Schatzmann *et al.*, 2007; Peng *et al.*, 2018; Li *et al.*, 2017; Kreyssig *et al.*, 2018) Malheureusement, des approches similaires ne peuvent pas être utilisées pour les tâches de recherche complexes et exploratoires (Belkin & Croft, 1992). Dans la RI à domaine ouvert, les facettes qui sous-tendent les besoins en information ne sont pas nécessairement discrètes ou facilement identifiables, ce qui rend beaucoup plus difficile l'identification et l'annotation des besoins des utilisateurs.

2.2 Question de clarification

Poser des questions de clarification est une tâche conversationnelle qui permet à l'utilisateur de participer au processus de désambiguïsation des requêtes en interagissant avec le système. Aliannejadi et al. (Aliannejadi *et al.*, 2019) proposent un classifieur qui sélectionne itérativement une question de clarification à chaque tour de conversation parmi un ensemble de questions prédéfini. Bi et al. (Bi *et al.*, 2021) complètent cette approche avec une détection d'intention basée sur les retours négatifs et un BERT basé sur la pertinence marginale maximale. Cependant, l'utilisation d'un ensemble de questions fixes limite la couverture des sujets, et donc l'efficacité de l'approche. Une deuxième ligne de travaux vise plutôt à générer des questions de clarification. Dans (Salle *et al.*, 2021), Salle et al. utilisent des modèles et des facettes collectés à partir de l'API Autosuggest Bing pour générer des questions de clarification. À chaque tour de la conversation, ils sélectionnent une nouvelle facette pour générer la question jusqu'à ce que la réponse de l'utilisateur soit positive. Sekulić et al (Sekulić *et al.*, 2021) proposent d'améliorer encore la fluidité en utilisant un LLM pour conditionner la génération de questions de clarification à la requête initiale et à une facette. Par exemple, la requête 'Tell me about kiwi', conditionnée aux facettes 'information fruit' ou 'biology birds' peut générer des questions telles que 'Are you interested in kiwi fruit ?' ou 'Are you interested in the biology of kiwi birds ? Ils s'appuient sur le jeu de données Clariq pour affiner GPT2, et ont constaté que les questions générées sont plus naturelles et utiles que les méthodes basées sur des modèles. Ils ont étendu ce travail en générant des questions à l'aide de facettes extraites des documents récupérés (Sekulić *et al.*, 2022). Zamani et al. (Zamani *et al.*, 2020a) proposent de générer des questions de clarification associées à de multiples facettes (panneaux de clarification) qui sont collectées à l'aide de données de reformulation de requêtes, les clics utilisateurs sont aussi collectés. En revanche, ces données sont construites seulement sur les reformulations de requêtes les plus probables et ne sont pas associées une collection ou des signaux de pertinence sur les documents.

Cet état de l'art met en évidence le manque de données à grande échelle adéquats contenant des interactions à initiatives mixtes pour la tâche de RI. Sachant que la collecte de ces jeux de données avec des annotations humaines serait coûteuse, nous pensons qu'une alternative possible est de générer automatiquement des interactions à initiative mixte à partir des collections existantes.

3 Simulation des interactions à initiatives mixtes et génération de jeux de données de RI conversationnelle

3.1 Définition du problème

Nous présentons notre méthodologie pour générer automatiquement des ensembles de données de RI à grande échelle et à initiative mixte. Pour ce faire, nous proposons d'augmenter les ensembles de données de RI ad hoc avec des interactions utilisateur-système simulées, à savoir des questions de clarification (pour le côté système) et les réponses correspondantes (pour le côté utilisateur).

Pour fournir un jeu de données utiles à l'entraînement de modèles neuronaux de recherche d'information avec des questions de clarification et des signaux de pertinence, il est important de fournir un large éventail d'interactions, à savoir des questions de clarification qui donnent lieu à des réponses positives ou négatives. En gardant à l'esprit qu'un sujet peut être complexe ou ambigu, nous suivons les travaux précédents (Sekulić *et al.*, 2021; Zamani *et al.*, 2020a; Salle *et al.*, 2021) en exploitant les facettes pour générer ces questions de clarification. L'extraction de facettes positives ou négatives autour d'un sujet peut être considérée comme un proxy pour limiter la génération de questions de clarification qui attendent des réponses par "oui" ou "non". De plus, pour assurer la qualité des interactions, nous proposons d'introduire une autre variable de contrainte modélisant l'intention de recherche de l'utilisateur. La paire des variables facette et intention permet de générer des questions de clarification positives et négatives (grâce à la facette) en gardant toujours la génération de réponses cohérentes avec les jugements de pertinence dans le jeu de données initial (grâce à l'intention). Autrement dit, l'échantillonnage de différentes paires facette-intention à partir de passages dont le jugement de pertinence est connu permet de constituer un jeu de données avec des interactions mixtes positives et négatives qui reflètent l'intention de recherche de l'utilisateur. Pour des raisons de simplicité, nous ne considérons que les interactions à un seul tour, et discutons de l'extension aux interactions à plusieurs tours dans la section 6.

Considérons un jeu de données de RI ad hoc $\mathcal{D} = \{\mathcal{P}, \mathcal{Q}, \mathcal{R}\}$, dans lequel $\mathcal{P}$ est une collection de passages (ou documents), $\mathcal{Q}$ est un ensemble de requêtes, et $\mathcal{R}$ est un ensemble de jugements de pertinence. L'ensemble $\mathcal{R}$ comprend des tuples $(q, \mathcal{P}_q^+, \mathcal{P}_q^-)$ indiquant les passages pertinents $\mathcal{P}_q^+ \subset \mathcal{P}$ et non pertinents $\mathcal{P}_q^- \subset \mathcal{P}$, pour une requête $q \in \mathcal{Q}$. Nous supposons que $\mathcal{P}_q^- \cap \mathcal{P}_q^+ = \varnothing$. Notre objectif est d'augmenter ce jeu de données $\mathcal{D}$ avec un ensemble d'interactions à initiative mixte $X = \{X_1, \ldots, X_i, \ldots, X_n\}$. Nous notons une interaction à initiative mixte $X_i = (q, cq, a)$ où q désigne une requête initiale, cq une question de clarification et a la réponse associée. Dans cette optique, nous concevons une méthodologie de construction de jeu de données $\mathcal{M} : D \cup \{X_1, \ldots, X_i, \ldots, X_n\}$ reposant sur deux étapes principales : 1) l'extraction des facettes (positives et négatives) f liées à chaque sujet (si elles ne sont pas disponibles dans le jeu de données initial de RI ad-hoc) qui est ensuite utilisé pour contraindre la génération de questions de clarification, et 2) la génération d'interactions à initiative mixte étant donné une requête q et cette facette f. Selon le jeu de données, les ensembles de facettes positives $\mathcal{F}^+$ et négatives $\mathcal{F}^-$ associés à la requête q peuvent être disponibles ou doivent être construits (section 3.2). Nous supposons également que l'intention de recherche int de l'utilisateur est caractérisée par les documents pertinents disponibles dans le jeu de données initial. Nous proposons ensuite de générer une interaction d'initiative mixte X_i étant donné une requête q et les variables de contrainte f et int. Nous nous appuyons sur 1) un modèle de clarification $\mathcal{CM} : q, f \rightarrow cq$ visant à générer une question de clarification cq autour de la facette f étant donné le sujet de la requête q, et 2) une simulation utilisateur $\mathcal{US} : (cq, int, f) \rightarrow a$ qui infère la réponse a en réponse à la question de clarification cq étant donné la pertinence de la facette f et l'intention de l'utilisateur int.
Nous présentons ci-dessous la méthode d'extraction des facettes, ainsi que les composants permettant de générer des questions et des réponses. Des exemples d'interactions attendues sont présentés dans le Tableau 1.

3.2 Extraction de Facettes

Les facettes peuvent être explicites ou implicites selon le jeu de données. Par exemple, elles sont spécifiées dans TREC Web 2009-12 (Clarke *et al.*, 2009), et par conséquent, Qulac et ClariQ (Over, 2001)). Si elles ne sont pas explicitement spécifiées, nous proposons de les extraire des documents. Des travaux antérieurs ont montré que les facettes des requêtes peuvent être extraites des documents

Exemple d'Interactions simulées	
Query : find phone number for irs **System** : are you looking for phone numbers for the irs internal revenue service ? **Answer** : yes **Intent** : Internal Revenue Service Phone Numbers. 800-829-1040 For individual and joint filers who need procedural or tax law information and/ or help to file their 1040-type IRS Tax Help Line for individual returns (including Individuals| Schedules C and E) ; and, general account information for Form 1040 Filers. **Facet** : internal revenue service phone numbers	**Query** : lps laws definition **System** : are you looking for a definition of a limited partnership ? **Answer** : no **Intent** : The Court will not let you establish an LPS conservatorship unless it finds beyond a reasonable doubt, that the mentally ill person, is gravely disabled. Gravely disabled means that, because of a mental disorder, the person cannot take care of his/her basic, personal needs for food, clothing, or shelter. **Facet** : limited partnership business

TABLE 1 – Exemples d'interactions simulées appartenant au jeu de données MiMarco. Dans le premier exemple, l'intention et la facette sont extraites d'un passage pertinent. Dans le deuxième exemple, l'intention est extraite d'un passage pertinent, mais la question de clarification fait référence à une facette négative du sujet.

les mieux classés (Dou *et al.*, 2016; Kong & Allan, 2013). Inspirés par l'analyse fournie par Sekulić et al. (Sekulić *et al.*, 2022), nous extrayons les mots-clés contextuels les plus importants pour représenter les facettes, comme suggéré dans (Sharma & Li, 2019). L'objectif de l'extraction de facettes est de fournir des mots-clés supplémentaires qui peuvent être utilisés pour générer ultérieurement une question de clarification sur divers sujets ou sous-thèmes. Dans ce travail, les facettes sont un ensemble de mots-clés fournissant un contexte supplémentaire à la requête. Nous la formulons comme une fonction bijective $\psi(P) :\rightarrow \mathcal{F}$ qui fait correspondre un ensemble P de passages à un ensemble de facettes. Étant donné une requête q, nous construisons les ensembles $\mathcal{F}^+$ et $\mathcal{F}^-$ de facettes positives et négatives à partir des ensembles de passages respectivement pertinents et non pertinents, respectivement $\mathcal{P}_q^+$ et $\mathcal{P}_q^-$. Cela nous permet de conserver la pertinence des facettes. Pour ce faire, pour un passage $p \in (\mathcal{P}_q^+ \cup \mathcal{P}_q^-)$, nous extrayons comme facette $f \in \mathcal{F}$ l'ensemble des K mots du passage qui sont les plus similaires avec la représentation du passage (c'est-à-dire avec le jeton [CLS]). Pour calculer la similarité, nous utilisons un modèle Sentence-Bert (MiniLM-L6-v2) pré-entraîné (Reimers & Gurevych, 2019) .

3.3 Génération des Interactions

3.3.1 Génération des questions de clarification

L'objectif du modèle de clarification $\mathcal{CM}$ est de poser des questions de clarification pertinentes relatives à une ambiguïté de la requête. Dans la plupart des modèles proposés (Zamani *et al.*, 2020a; Sekulić *et al.*, 2022; Sekulić *et al.*, 2021; Salle *et al.*, 2021; Aliannejadi *et al.*, 2019), cette ambiguïté est traitée en utilisant le concept de facette. Ainsi, la génération de questions de clarification cq est conditionnée par la requête initiale q et une facette f :

$$p(cq|q, f) = \Pi_i p(cq_i|cq_{<i}, q, f) \tag{1}$$

ou q_i est le i^e jeton de la séquence et $q_{<i}$ les jetons décodés.

3.3.2 Simulation Utilisateur

L'objectif de la simulation utilisateur $\mathcal{US}$ est d'imiter la réponse de l'utilisateur en réponse à une question de clarification compte tenu de son intention. La simulation utilisateur doit donner des réponses utiles aux questions pour aider le système à comprendre son intention. L'intention est une représentation du besoin d'information liée à la requête initiale. Elle est utilisée pour orienter la réponse de la simulation utilisateur vers cet objectif. (Kang *et al.*, 2019b; Gao *et al.*, 2022; Wu *et al.*, 2020; Zhou *et al.*, 2020b; Fu *et al.*, 2020; Erbacher *et al.*, 2022). Nous limitons la question de clarification à demander si l'intention porte sur une facette et la réponse de la simulation d'utilisateur à "oui" ou "non". Cette forme limitée de réponse est motivée par deux raisons : (1) malgré la simplicité, une réponse correcte de cette forme correspond aux interactions réalistes de base avec les utilisateurs et est très utile pour que le système puisse mieux identifier l'intention derrière la requête. (2) Cette forme simple de question et de réponse est plus facile à générer et à évaluer. Plus formellement, la simulation de l'utilisateur vise à estimer la probabilité d'une réponse $a \in \{yes, no\}$ étant donné une requête q, une intention de recherche int, et une question de clarification : $p(a|q, int, cq)$.

Intention de recherche.L'intention de l'utilisateur correspond au besoin d'information de l'utilisateur et n'est connue que par ce dernier. Bien que de multiples représentations de l'intention puissent être adoptées (comme une description détaillée du besoin d'information (Aliannejadi *et al.*, 2019, 2021), une représentation vectorielle (Erbacher *et al.*, 2022) ou des contraintes (Kang *et al.*, 2019b; Gao *et al.*, 2022; Wu *et al.*, 2020; Zhou *et al.*, 2020b; Fu *et al.*, 2020)), les jeux de données de RI n'ont généralement pas d'intention annotée associée à la requête. Cependant, les passages pertinents sont connus dans un jeu de données de RI. Dans cet article, nous utilisons un passage pertinent échantillonné $p \in \mathcal{P}_q^+$ et assimilons son contenu à l'intention sous-jacente int. Formellement : $int \leftarrow p$. Nous reconnaissons que ce choix repose sur une hypothèse forte et nous en discutons dans la section 7.

3.4 Utilisation des interactions d'initiative mixte pour adapter les jeux de données RI ad hoc à la RI conversationnelle

Étant donné un jeu de données de RI ad hoc $\mathcal{D}$, notre objectif est d'augmenter $\mathcal{D}$ avec des conversations d'initiative mixte X. Nous distinguons la création des jeux de données d'entraînement et de test, car ils ont des objectifs différents. Les données d'entraînement nécessitent d'inclure des interactions positives et négatives pour permettre d'entraîner correctement les modèles neuronaux orientés RI face à divers scénarios. Pour rappel, ces interactions positives/négatives sont construites sur la base de documents pertinents et non pertinents déterminant des facettes positives et négatives. L'utilisation de la même heuristique pour générer un jeu de données de test n'est pas appropriée car cela impliquerait d'inclure les jugements de pertinence comme sources de preuves de la génération de questions de clarification à l'étape d'inférence. Par conséquent, nous proposons de concevoir une méthodologie d'évaluation en ligne, exploitant le modèle de clarification $\mathcal{CM}$ et la simulation utilisateur $\mathcal{US}$ pour générer des interactions sans introduire de biais lié aux jugements de pertinence. Nous présentons ces deux méthodologies visant à générer des jeux de données hors ligne et en ligne dans ce qui suit.

3.4.1 Construction un jeu d'entraînement hors ligne avec des jugements de pertinence

Notre méthodologie hors ligne vise à générer un large éventail d'interactions positives et négatives sur la base d'un jeu de données RI ad-hoc. Pour ce faire, nous utilisons des documents pertinents et non pertinents pour construire des facettes positives et négatives contraignant la génération de questions de clarification. Comme contrainte de qualité supplémentaire dans la supervision du jeu de données, nous souhaitons nous assurer que les réponses correspondent à la pertinence des documents utilisés.

Autrement dit, la simulation utilisateur présentée dans la section 3.3.2 est remplacée par une simple heuristique faisant correspondre les réponses a avec la pertinence des facettes f (équation 2 page suivante).

Require: $\mathcal{D} = \{\mathcal{P}, \mathcal{Q}, \mathcal{R}\}$
  $X \leftarrow \{\}$ ▷ Ensemble d'interactions RI
  for $q \in \mathcal{Q}$ **do**
    $\mathcal{F}^+ \leftarrow \psi(\mathcal{P}_q^+)$ ▷ Extraction facets positive
    $\mathcal{F}^- \leftarrow \psi(\mathcal{P}_q^-)$ ▷ Extraction facets négative
    for $f \in (\mathcal{F}^+ \cup \mathcal{F}^-)$ **do**
      $cq \leftarrow \mathcal{CM}(q, f)$ ▷ Génération de questions
      if $f \in \mathcal{F}_q^+$ **then** ▷ Construction réponse
        $a \leftarrow$ *'yes'*
      else
        $a \leftarrow$ *'no'*
      end if
      $X_i = (q, cq, a)$
      $X \leftarrow X \uplus X_i$
    end for
  end for
  return $\mathcal{D} \cup X$

Algorithm 1 – Méthodologie pour construire un jeu de donnée d'entraînement à initiative mixé pour la RI

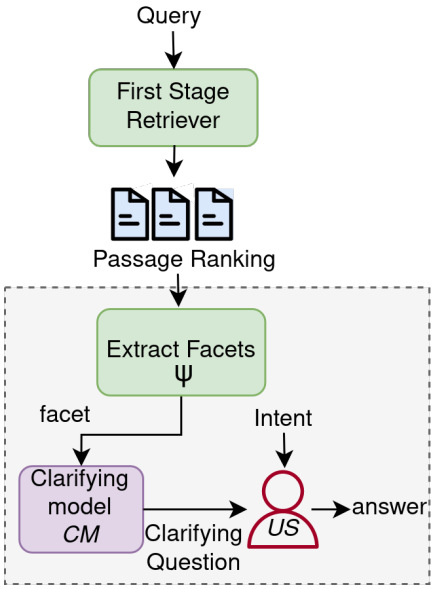

FIGURE 1 – Méthodologie d'évaluation en ligne pour créer des interactions à initiative mixte sur le jeu de test

$$a = \begin{cases} 'yes' \; si \; f \in \mathcal{F}^+ \\ 'no' \; sinon \end{cases} \tag{2}$$

Nous proposons une méthodologie en 3 étapes présentée dans l'Algorithme 1. Étant donné une requête q : 1) des facettes positives et négatives, respectivement $\mathcal{F}^+$ et $\mathcal{F}^-$, sont extraites des ensembles de passages pertinents et non pertinents, respectivement $\mathcal{P}_q^+$ et $\mathcal{P}_q^-$; 2) une interaction X_i est émise pour une facette f, générant la question de clarification associée cq (avec $\mathcal{CM}$) et associant la réponse a à la pertinence de la facette (équation 2) ; 3) l'ensemble d'interactions X est incrémenté avec cette nouvelle interaction X_i, ce qui permet de construire un jeu de données de RI à initiative mixte en associant l'ensemble d'interactions X construit sur toutes les requêtes du jeu de données de RI ad hoc initial $\mathcal{D}$.

3.4.2 Construction des données de test pour l'évaluation en ligne sans jugement de pertinence

Notre méthodologie en ligne vise à générer des interactions sans s'appuyer sur la pertinence des documents. Au lieu de cela, nous tirons parti de la rétroaction de pseudo-pertinence en utilisant les SERPs d'un modèle de recherche d'information comme un proxy pour extraire les facettes de la requête. Chaque facette conditionne la génération de la question de clarification et de la réponse. Plus particulièrement, le pipeline proposé pour générer des interactions en ligne pour une requête q est présenté dans la figure 1. Il est construit sur les étapes suivantes : 1) classement des documents à l'aide d'un modèle de recherche d'information (dans notre cas BM25), 2) extraction de l'ensemble des facettes sur la base des documents pseudo-pertinent, et 3) génération de l'interaction.

Selon les besoins de l'évaluation, différents choix peuvent être faits concernant l'extraction des

facettes. On peut extraire une seule facette du document le mieux classé afin d'effectuer une seule étape de recherche pour une requête (la stratégie utilisée dans nos expériences). D'autres tâches ou objectifs d'évaluation nécessiteraient la génération de multiples facettes et, par conséquent, de multiples clarifications. Cela peut être fait en identifiant les documents les plus hauts/les plus bas obtenus avec le classement de la première étape comme des documents pseudo-pertinents ; chaque document conditionnant l'extraction de facettes comme décrit dans la section 3.2.

4 Évaluation de la méthodologie de génération

Dans cette section, nous évaluons notre méthodologie, et en particulier, la qualité des interactions simulées. Veuillez noter que nous nous concentrons sur l'augmentation du jeu de données MsMarco, mais que notre méthodologie peut être généralisée à tout jeu de données de RI ad hoc.

4.1 Protocole d'Evaluation

4.1.1 Jeux de Données

Nous nous concentrons ici sur le jeu de données MsMarco 2021 passages (Nguyen *et al.*, 2016) qui est un jeu de données de RI à domaine ouvert contenant 8,8 millions de passages et plus de 500 000 paires de pertinence requête-passage avec environ 1,1 passage pertinent par requête en moyenne. MsMarco est couramment utilisé pour entrainer et évaluer les architectures de recherche d'information (Thakur *et al.*, 2021). Nous tirons parti des passages de MsMarco pour extraire des exemples négatifs et donc nos tuples $(q, \mathcal{P}^+, \mathcal{P}^-)$ en s'appuyant sur des modèles de l'état de l'art (Reimers & Gurevych, 2019) [1]. Pour entraîner le modèle de clarification $\mathcal{CM}$, nous utilisons la version filtrée du jeu de données ClariQ proposé dans (Sekulić *et al.*, 2021) qui associe des questions de clarification à des facettes. Toutes les questions de clarification de ce jeu de données sont construites de manière à attendre des réponses " oui " ou " non ". Ce jeu de données fournit 1756 tuples supervisés de (requête-facette-question clarificatrice) pour 187 requêtes.

Pour entraîner la simulation utilisateur $\mathcal{US}$, nous exploitons la moitié de l'ensemble d'entraînement du jeu de données MsMarco (250000 requêtes) pour extraire les facettes positives et négatives comme détaillé dans la section 3.2 et générer des questions de clarification en utilisant le modèle $\mathcal{CM}$. L'étiquette de supervision liée aux réponses est déduite comme proposé dans l'évaluation hors ligne (voir l'équation 2).

Pour l'évaluation hors ligne, étant donné que le jeu de données original comprend des annotations éparses, c'est-à-dire que certains passages sont pertinents mais ne sont pas annotés en tant que tels, il est possible que des documents pertinents soient considérés comme des documents non pertinents. Cette tendance se retrouve toutefois dans l'ensemble d'entraînement MsMarco qui ne comprend qu'un seul document pertinent par requête. Par conséquent, pour assurer la cohérence de l'étiquetage, nous suivons (Qu *et al.*, 2021) et débruitons les exemples négatifs dans l'ensemble du jeu d'entraînement en utilisant un modèle d'encodeur croisé pré-entraîné [2] qui capture les similarités entre les passages. Pour l'évaluation en ligne, nous avons choisi de générer une seule interaction basée sur le document le mieux classé afin de correspondre à notre tâche d'évaluation extrinsèque basée sur la RI. Nous publierons, après acceptation, les jeux de données générés complets ainsi que le modèle de clarification $\mathcal{CM}$ et la simulation utilisateur $\mathcal{US}$ pour permettre la génération d'interactions supplémentaires. Le tableau 1 présente quelques exemples de conversations simulées générées à partir de requêtes MsMarco.

1. https ://huggingface.co/datasets/sentence-transformers/msmarco-hard-negatives
2. https://huggingface.co/cross-encoder/ms-marco-MiniLM-L-6-v2

4.1.2 Modèles de référence et métriques

Évaluation des questions de clarification avec des métriques automatiques. Nous nous appuyons sur (Sekulić *et al.*, 2021) et comparons notre modèle de clarification, noté $\mathcal{CM}$, avec 1) une approche basée sur un modèle ($Template$). Le modèle suit une séquence prédéfinie concaténant des facettes : 'Are you looking for + Facet'. 2) $\mathcal{CM}w/oFacet$: la version de notre modèle $\mathcal{CM}$ uniquement conditionnée par la requête. Il s'agit en fait d'un modèle T5 entraîné comme modèle de traduction automatique, qui génère une question de clarification à partir de la requête uniquement. Nous évaluons la capacité de $\mathcal{CM}$ à générer des questions de clarification en utilisant les références fournies dans l'ensemble de test ClariQ. Nous prenons en compte la métrique METEOR (Banerjee & Lavie, 2005) et la similarité moyenne en cosinus entre les phrases incorporées (COSIM). METEOR est couramment utilisé pour évaluer les résultats de la traduction automatique en considérant le rappel et la précision unigramme. Au niveau de la phrase, cette méthode présente une bonne corrélation avec les jugements humains (Banerjee & Lavie, 2005). Pour calculer le score de similarité, nous encodons les questions à l'aide d'un transformer MiniLM-L6-v2 bien entraîné (Reimers & Gurevych, 2019). Nous utilisons le t-test pour évaluer la significativité des différences (*** : valeur p<0,005). Pour évaluer si les questions générées sur MsMarco sont similaires à leur passage relatif, nous calculons également la similarité cosinus moyenne entre les questions de clarification et leurs passages pertinents et non pertinents récupérés. Nous encodons les questions en utilisant MiniLM-L6-v2 (Reimers & Gurevych, 2019).

Évaluations humaines sur des questions de clarification. Pour comparer et mieux évaluer la qualité d'une question de clarification générée sur MsMarco, nous avons effectué une évaluation humaine. À partir de la requête initiale de l'utilisateur et du passage utilisé pour générer la question, nous avons demandé aux annotateurs d'évaluer la qualité de 200 questions de clarification échantillonnées parmi les trois modèles ($Template$, $\mathcal{CM}w/oFacets$ et notre modèle $\mathcal{CM}$). Pour ce faire, les annotateurs sont invités à sélectionner une question de clarification préférée parmi les trois suggestions affichées dans un ordre mélangé pour les critères suivants : **1) L'Utilité** qui évalue si une question peut aider à mieux comprendre ou à affiner la requête en fournissant des informations ou des suggestions supplémentaires. **2) Le Naturel** qui évalue la fluidité et la lisibilité de la question. **3) La Pertinence** qui évalue si une question est spécifique ou liée aux informations contenues dans un passage. Chaque annotateur a évalué 20 cas différents et, pour chaque métrique, a identifié le meilleur modèle de sortie. Nous avons recruté 10 évaluateurs. Chaque instance est évaluée par 2 annotateurs et nous obtenons une métrique Kappa égale à 0,324 montrant un accord équitable entre les évaluateurs. Nous avons également distingué les résultats pour les réponses positives et négatives des utilisateurs en échantillonnant les facettes pertinentes et non pertinentes.

Évaluations humaines sur les réponses. Une hypothèse forte de notre méthode est que les questions de clarification générées avec des facettes extraites de passages pertinents conduisent à des réponses positives alors que l'utilisation de passages non pertinents pour générer des facettes négatives conduit intrinsèquement à des réponses négatives. Pour valider cette hypothèse forte, nous avons montré à des évaluateurs humains différentes instances comprenant une requête q, une question de clarification cq, et le passage pertinent p utilisé pour construire la facette f. Pour chaque instance, nous avons demandé aux évaluateurs humains de répondre par 'oui' ou 'non' aux questions de clarification. Cette évaluation humaine implique 10 annotateurs humains pour un total de 200 questions, avec un équilibre entre les facettes pertinentes et non pertinentes utilisées pour générer la question de clarification. Chaque instance est annotée par deux humains. Nous obtenons une métrique de Kappa égale à 0,472 montrant un accord modéré entre les évaluateurs. Pour valider notre hypothèse, nous considérons les

réponses humaines comme référence et nous les comparons à notre méthode d'étiquetage automatique (à savoir, la simulation utilisateur $\mathcal{US}$) pour calculer la métrique de précision.

4.1.3 Détails d'implémentation

Pour les deux modèles $\mathcal{CM}$ et $\mathcal{US}$, nous avons utilisé le point de contrôle T5 pré-entraîné disponible sur le hub Huggingface (Raffel *et al.*, 2020; Wolf *et al.*, 2019). Pour affiner ces deux modèles, nous avons utilisé le Teacher Forcing (Williams & Zipser, 1989) , nous utilisons AdaFactor (Shazeer & Stern, 2018), et un taux d'apprentissage de 5.10^{-5} avec des tailles de batch de 64. L'incorporation des mots est calculée à l'aide d'un modèle MiniLM-L6-v2 pré-entraîné (Reimers & Gurevych, 2019). Le nombre de mots extraits est fixé à $k = 5$ pour l'ensemble des expériences. Pour l'inférence, nous utilisons l'échantillonnage du noyau (p=0,95) pour les modèles $\mathcal{CM}$ et $\mathcal{US}$.

4.2 Évaluation des interactions générées

4.2.1 Évaluation Automatique

Le tableau 2 rapporte l'efficacité du modèle de clarification sur l'ensemble de test ClariQ. Les résultats montrent que notre modèle $\mathcal{CM}$ surpasse significativement tous les modèles de référence. Les résultats inférieurs obtenus par le modèle de référence $\mathcal{CM}w/oFacet$ soulignent que modèle sans facette est moins efficace que les modèles utilisant des facettes. Les facettes sont utiles pour contraindre le modèle de clarification, et les modèles seq-to-seq basés sur de grands modèles de langage sont plus naturels que les méthodes basées sur des templates. Les facettes sont extraites d'un passage pertinent ou non pertinent et utilisées pour générer des questions de clarification. Le tableau 3 rapporte la similarité en cosinus entre les questions encodées et les passages respectifs. Nous observons que la similarité entre les questions de clarification et leurs passages connexes (en gras) est plus élevée que celle entre les questions de clarification et les requêtes. Cela montre que les questions générées ne sont pas génériques à la requête mais orientées vers les passages fournis.

	METEOR	COSIM
$Template$	0.338^{***}	0.643^{***}
$\mathcal{CM}w/oFacet$	0.326^{***}	0.608^{***}
$\mathcal{CM}$	**0.557**	**0.812**

TABLE 2 – Évaluation des questions de clarification sur le test de ClariQ. *** pour les résultats significatifs avec le modèle $\mathcal{CM}$ (p<0.005)

	q	p+	p-
cq+	0.675	**0.721**	0.450
cq-	0.521	0.450	**0.685**

TABLE 3 – Similarité cosinus entre les questions et leur passage respectif. $cq+$, $cq-$ pour les questions positives et negatives.

4.2.2 Évaluation Humaine

Nous présentons l'évaluation humaine des questions de clarification dans le tableau 4. Le modèle $\mathcal{CM}$ sans facette génère des questions plus naturelles que les autres modèles (préféré pour 46.3% de l'échantillon). Le modèle $\mathcal{CM}$ ajusté avec facette génère plus de questions utiles et pertinentes, ce modèle est considéré comme le plus pertinent par les évaluateurs dans $59,9\%$ de l'échantillon testé. Cela montre que la facette récupérée dans la génération aide à générer des questions plus utiles et plus pertinentes.

Dans l'évaluation humaine des réponses, nous obtenons une précision de 0,685 entre les réponses humaines et l'étiquetage automatique des questions de clarification. Il existe de multiples causes expliquant la différence entre les réponses humaines et l'étiquetage automatique. 1) Facette ne capture pas toujours correctement les informations fournies dans un passage, ce qui conduit à des questions

	Answer	Naturalness	Usefulness	Relevance
Template	positive	0.044	0.086	0.120
	negative	0.073	0.095	0.146
	total	0.119	0.181	0.267
$\mathcal{CM}w/oFacet$	positive	**0.243**	0.195	0.077
	negative	**0.220**	0.140	0.056
	total	**0.463**	0.336	0.133
$\mathcal{CM}$	positive	0.206	**0.213**	**0.297**
	negative	0.211	**0.268**	**0.301**
	total	0.417	**0.481**	**0.599**

TABLE 4 – Résultats de l'évaluation humaine sur Msmarco-passage. Le $\mathcal{CM}$ sans facette produit des questions plus naturelles, mais pas aussi pertinentes que le $\mathcal{CM}$ avec facette.

de clarification de mauvaise qualité. 2) Le modèle $\mathcal{CM}$ ne génère pas toujours une question orientée vers la facette fournie et produit une reformulation de la requête initiale, posant ainsi une question non liée à une facette.

5 Évaluation sur une tâche de RI

Dans cette section, nous proposons d'évaluer indirectement la qualité du jeu de données généré à travers une tâche de RI. En effet, des travaux précédents (Qu *et al.*, 2020; Zhou *et al.*, 2020b; Li *et al.*, 2018a; Fu *et al.*, 2020; Jia *et al.*, 2022) ont déjà utilisé des tâches extrinsèques pour valider un jeu de données. Par conséquent, nous introduisons un modèle de recherche d'information neuronal qui estime les scores de pertinence des passages en fonction de la requête et d'une interaction d'initiative mixte. Notre objectif est double : 1) L'application de ce modèle à notre jeu de données généré permet de savoir si la question de clarification et la réponse associée donnent effectivement un retour utile pour mieux comprendre le besoin d'information sous-jacent. L'évaluation est basée sur l'hypothèse suivante : si un modèle de RI utilisant les interactions générées est plus performant que celui qui ne les utilise pas, les interactions sont considérées comme pertinentes et utiles. 2) Nous fournissons une première base de modèles de référence pour les tâches de RI à initiative mixte.

5.1 Modèle de RI tirant parti des interactions d'initiative mixte

Nous proposons un modèle simple basé sur une architecture d'encodeurs croisés qui s'est avéré efficace pour les tâches de RI, en particulier lors de l'utilisation de modèles de langage de grande taille : (Pradeep *et al.*, 2021). L'encodeur croisé précédent vise à prédire la pertinence d'un passage p étant donné une requête q $P(relevant = 1|q, p)$. Notre modèle estime un score pour les passages en fonction de la requête, d'une question de clarification et d'une réponse de l'utilisateur (q, cq, a) :

$$p(relevant = 1|p, q, cq, a) \tag{3}$$

Suivant (Pradeep *et al.*, 2021), le score de pertinence est calculé grâce la log-probabilité prédite des tokens vrai/faux :

$$s_p = \log p(true|q, p, cq, a) \tag{4}$$

Suivant (Pradeep *et al.*, 2021), nous utilisons le modèle MonoT5 et intégrons des interactions d'initiative mixte en plus de la requête initiale pour mieux d'estimer les scores des documents. La séquence d'entrée est une concaténation de la requête, du document, de la question et de la réponse, séparés par des tokens spéciaux :

$$\text{Query : q Document : d Question : cq Answer : a} \tag{5}$$

	MRR@10	NDCG@1	NDCG@3	NDCG@10
BM25	0.1840***	0.105***	0.1690***	0.228***
BM25 + RM3	0.1566***	0.0807***	0.1386***	0.2021***
BM25 + MonoT5	0.3522***	0.2398***	0.3457***	0.4034***
BM25 + CLART5	**0.3863**	**0.2788**	**0.3817**	**0.4327**

TABLE 5 – Performance de RI sur MiMarco test. *** : two-sided t-test w.r.t. BM25+CLART5. with p-value<0.005

5.2 Détails d'entraînement

Nous avons utilisé MonoT5 pré-entraîné disponible sur le hub Huggingface (Raffel *et al.*, 2020; Wolf *et al.*, 2019). Nous affinons ce modèle sur notre ensemble d'entraînement, en utilisant teacher forcing et entropie croisée. Nous considérons une longueur de séquence maximale de 512 et une taille de batch de 128 séquences. Afin d'apprendre correctement à faire la distinction entre les passages pertinents et non pertinents d'une question, nous intégrons les interactions négatives dans le batch.

Pour l'optimisation, nous utilisons AdaFactor (Shazeer & Stern, 2018), et un taux d'apprentissage de 10^{-4}. Le réglage fin du modèle prend environ 4 heures sur 4 RTX 3080 (24 Go).

Au moment du test, nous effectuons une recherche de documents sur la requête initiale en utilisant l'implémentation pyserini (Lin *et al.*, 2021) de BM25. Nous appliquons ensuite notre modèle comme un modèle d'ordonnancement avec des informations supplémentaires. Nous fixons le nombre de documents récupérés à 100.

5.3 Métriques et modèles de référence

Nous utilisons des mesures classiques pour évaluer la qualité de l'ordonnancement des documents, à savoir le gain cumulé actualisé normalisé (NDCG) aux rangs 1, 3 et 10, et le rang réciproque moyen (MRR) au rang 10. Pour évaluer le potentiel de notre jeu de données, nous comparons les performances de notre modèle, noté **BM25+CLART5**, aux approches suivantes :

— **BM25.** BM25 est un modèle d'ordonnancement connu qui s'appuie sur la fréquence des mots contenu dans les documents, couramment utilisé comme référence.

— **BM25 + RM3.** RM3 est une méthode de pseudo-pertinence pour l'expansion de requête. La requête est développée à l'aide de termes d'expansion extraits des 10 premiers documents récupérés. RM3 est une base de référence compétitive et est souvent utilisée pour évaluer les modèles de RI. (Thakur *et al.*, 2021; Adolphs *et al.*, 2022).

— **BM25 + MonoT5.** MonoT5 est un modèle d'ordonnancement pré-entraîné sur l'ensemble d'entraînement original de MsMarco, c'est-à-dire uniquement les requêtes et les jugements de pertinence. Ce modèle atteint des performances de pointe sur le tableau de classement beir (Thakur *et al.*, 2021) et constitue une référence naturelle puisque BM25+CLART5 utilise le même modèle pré-entraîné de deuxième étape avant de l'affiner sur des interactions.

5.4 Efficacité de l'ordonnancement neuronal orienté initiative mixte

Nous présentons les résultats de notre modèle d'ordonnancement neuronal à initiative mixte obtenus sur le pipeline d'évaluation en ligne présenté dans la section 3.4 appliqué sur l'ensemble de test MsMarco (Tableau 5).

Le tableau 5 met en évidence le fait que les informations supplémentaires permettent à BM25+CLART5 d'améliorer de manière significative toutes les métriques sur le jeu de données MsMarco augmenté. Par exemple, BM25+CLART5 augmente le score MRR@10 de 0,034 point par

rapport à BM25+MonoT5. Une analyse plus poussée des résultats sur MsMarco montre que pour 33.0% des requêtes, le passage pertinent n'est pas récupéré dans le top-100 par BM25, conduisant le MRR@100 à 0.0. Pour 25,6 des requêtes, Monot5 et ClarT5 obtiennent la même valeur MRR@10. Pour $30,3\%$des requêtes, BM25+CLART5 obtient un meilleur MRR@10 tandis que $11,1\%$ obtiennent un MRR@10 inférieur. Dans l'ensemble, ces résultats montrent que le feedback fourni par la simulation de l'utilisateur à la question de clarification est pertinent et utile. Il permet d'augmenter le classement des passages pertinents. Ce résultat confirme indirectement que les interactions simulées encodent effectivement des informations pertinentes pour les intentions de recherche sous-jacentes, ce qui correspond à ce que les utilisateurs réels fourniraient dans les conversations. Par conséquent, les simulations proposées sont raisonnables.

6 Expériences complémentaires

6.1 Extension aux interactions multi-tours

Dans la section précédente, nous avons simulé une interaction avec une seule requête $X = (q, cq, a)$ pour l'inférence en ligne. Cependant, de multiples facettes différentes peuvent être extraites des passages récupérés. Cela signifie que des séquences d'interactions $X_0, ..., X_t$ peuvent être inférées en sélectionnant séquentiellement différentes facettes. Bien qu'un nouvel ensemble de passages puisse être récupéré en utilisant la dernière interaction, nous ne considérons ici que les facettes des passages récupérés avec la requête initiale. Chaque t^{ieme} tour exploite le t^{ieme} document dans la liste de documents pour construire une facette et générer une question de clarification. Les interactions multi-tours sont donc générées dans un ordre non arbitraire.

Impact sur la conception du modèle de RI neuronal. Nous proposons d'étendre le modèle au re-classement multi-tour en utilisant plusieurs tours de clarification autour de la même requête. Nous évaluons les passages en utilisant des interactions multiples autour de la même intention de recherche. À chaque pas de temps t, un nouveau score s_d^t est calculé pour les passages du même classement utilisant une seule interaction. Ce score est calculé en utilisant l'équation 6 qui prédit les scores de pertinence cumulatifs à toutes les interactions, c'est-à-dire la somme des scores de pertinence jusqu'au temps T. Ce score est utilisé comme le score d'un document suivant une séquence d'interactions $X_t = \{q, cq^1, a^1, ..., cq^t, a^t\}$. cq^t et a^t sont la question de clarification et la réponse générées à l'instant t.

$$s_d^T = \sum_{t=0}^{T} \log p(relevant = 1 | q, p, cq^t, a^t) \tag{6}$$

où s_d^T est le score du document p au temps T. Comme le classement est mis à jour entre les tours, nous sélectionnons les facettes du passage le mieux classé à chaque pas de temps. Nous évaluons la performance d'ordonnancement à différentes longueurs d'interactions, de $T = 1$ à $T = 5$. Nous présentons également l'entropie du classement (Shannon, 1948) comme une mesure de la confiance du système en mesurant comment les scores sont distribués dans le classement. Cette entropie est maximisée lorsque la distribution des scores est uniforme sur le classement.

Resultats. Le tableau 6 présente les performances en fonction du nombre de tours de clarification pour le jeu de test MsMarco. Nous observons que l'augmentation du nombre de tours améliore les scores MRR, NDCG et tout en minimisant l'entropie, ce qui suggère que les informations supplémentaires aident à affiner davantage les besoins d'informations de l'utilisateur.

	MRR@10	NDCG@1	NDCG@3	NDCG@10	Entropy
BM25 + CLART5 T=1	0.3863	0.2788	0.3817	0.4327	2.951
BM25 + CLART5 T=2	0.44467	0.35186	0.43734	0.48038	2.303
BM25 + CLART5 T=3	0.48176	0.39828	0.47483	0.51089	2.163
BM25 + CLART5 T=4	0.50861	0.43266	0.50321	0.53183	2.06
BM25 + CLART5 T=5	**0.52949**	**0.45946**	**0.52311**	**0.54814**	**1.99**

TABLE 6 – Résultats d'efficacité sur les tours d'interaction T émis sur les ensembles de test MsMarco.

Multi-turn Conversations	**mrr@100**	**Entropy**
User : where was the bauhaus built	0.04167	3.854
System : do you want to know where the bauhaus was founded ? **User** : no	0.25	3.529
System : do you want to know where the gropius sommerfeld house was built ? **User** : yes	0.5	0.693
System : do you want to know where the bauhaus art school was built ? **User** : no	1.0	0.000218
User : what are tor	0.0	3.776
System : do you want to know if tor is free ? **User** : no	0.0	3.612
System : do you want to know more about tors transoral robotic surgery ? **User** : no	0.0	3.586
System : do you want to know how to connect to a computer network **User** : no	0.0	3.567
System : are you looking for a web browser for tor onion ? **User** : no	0.0	3.541
System : do you want to know about tor anonymity ? **User** : no	0.0	3.329

TABLE 7 – Exemples de cas réussis (en haut) et de cas d'échec (en bas). Dans le cas du haut, les questions de clarification aident à affiner l'intention de l'utilisateur et à mieux identifier les passages pertinents. Cet exemple montre une ambiguïté typique dans le document récupéré : le Bauhaus peut faire référence au groupe de rock, à l'école d'art, au prix... Pour le cas d'échec, aucun des passages pertinents n'est récupéré, posant des questions de clarification et reclassant les passages n'améliore pas les scores.

Nous décrivons des exemples qualitatifs de séquences de clarification réussies et non réussiesdans la table 7. Nous pouvons voir dans le premier exemple qu'une interaction supplémentaire permet de mieux affiner les scores du passage pertinent conduisant à un meilleur MRR@100, tandis que l'entropie diminue. Dans la dernière interaction, l'entropie est très faible, ce qui signifie que la distribution des scores est dense sur quelques passages. D'autre part, le deuxième exemple est un cas d'échec où les passages pertinents ne sont même pas récupérés. Dans certain cas d'échec que nous observons, où l'interaction tourne détériorer le classement, ce qui montre que les interactions générées ne sont pas toujours parfaites.

7 Conclusion et discussion

Il existe un besoin critique d'ensembles de données adéquats avec des interactions à initiative mixte pour la RI conversationnelle, mais la création d'un tel jeu de données est très coûteuse. La collecte à grande échelle de donnée de recherche conversationnelle interactive mixte dans un domaine ouvert avec un jugement de pertinence de document annoté reste très coûteuse. Dans cet article, nous avons proposé une méthode pour augmenter les ensembles de données IR ad hoc en simulant une forme simple d'interactions de clarification entre un utilisateur et un système. Cette méthode génère automatiquement les questions et les réponses à partir d'un grand jeu de données RI, permettant

d'expérimenter des approches RI conversationnelles à grande échelle. L'approche proposée est générique et peut être appliquée à tout jeu de données RI ad hoc existant. Dans les expériences, nous avons augmenté le jeu de données MsMarco et évalué la qualité des interactions avec les tâches intrinsèques et extrinsèques, en nous appuyant sur des métriques automatiques et des évaluations humaines. Les résultats montrent que, malgré la simplicité de nos approches, les interactions générées sont pertinentes pour les intentions de recherche et utiles pour un meilleur classement des documents. De plus, nous étendons notre méthodologie à un cadre de clarification multi-tours et fournissons des expériences préliminaires mettant en évidence le potentiel de notre méthodologie. Il s'agit d'une première approche pour l'augmentation d'ensembles de données à grande échelle pour la RI conversationnelle. Il démontre la faisabilité de la construction automatique de jeux de données. En tant que première investigation, cette étude présente plusieurs limites qui pourront être améliorées dans le futur.

— Dans un premier temps, notre investigation se limite à clarifier des questions basées sur une seule facette, souvent assimilées à des questions du type : "Fais-tu référence à 'facette' ?". Cependant, de véritables questions de clarification peuvent également poser des questions sur plusieurs sujets/facettes en un seul tour (ex : Êtes-vous intéressé à connaître *sujet1*, *sujet2* ou *sujet3*) ou également être formulées comme ouvertes questions terminées (par exemple, "Qu'aimeriez-vous savoir sur *sujet* ?"). Ces questions plus complexes sont plus difficiles à générer et à répondre dans les simulations, mais peut potentiellement apporter plus d'informations et être plus naturel dans la conversation.
— Deuxièmement, l'extraction des facettes reposait sur quelques mots-clés et cela peut être amélioré. Nous observons que lorsque les passages sont longs et traitent de plusieurs sujets, la question générée peut ne pas représenter le sujet abordé dans le passage. L'extraction des facettes doit être améliorée.
— Troisièmement, la simulation de l'utilisateur a été limitée aux réponses 'oui'/'non'. Dans une recherche conversationnelle plus sophistiquée, l'utilisateur pourrait fournir des informations plus et diverses dans la réponse. Simuler des réponses d'utilisateurs plus complexes est un défi pour l'avenir.
— Enfin, nous avons également généré des interactions multi-tours mais n'avons pas considéré la dépendance entre les tours. Dans une recherche conversationnelle réelle, les tours ultérieurs peuvent dépendre des précédents. Des simulations plus raisonnables d'interactions multitours devraient tenir compte de la dépendance. Ce cadre doit être profondément réfléchi. En effet, les conversations générées ne sont qu'une concaténation de plusieurs tours de clarification indépendants autour d'une même requête utilisateur. Il est crucial de définir une stratégie pour sélectionner la bonne séquence de questions de clarification afin d'optimiser la réussite de la session de recherche.

Malgré les limites, la démonstration de faisabilité faite dans cet article pour créer des jeux de données de RI conversationnelles à grande échelle ouvre la porte à d'autres enquêtes à grande échelle sur le sujet. Cela dit, nous espérons cependant que notre méthodologie aiderait la communauté à définir des cadres d'évaluation pour la RI conversationnelle en tirant parti des jeux de données RI ad hoc existants.

Remerciements

Nous tenons à remercier le projet ANR JCJC SESAMS (Projet-ANR-18-CE23-0001) pour son soutien à Pierre Erbacher et Laure Soulier de Sorbonne Université dans le cadre de ce travail.

Références

ADOLPHS L., HUEBSCHER M. C., BUCK C., GIRGIN S., BACHEM O., CIARAMITA M. & HOFMANN T. (2022). Decoding a neural retriever's latent space for query suggestion. DOI : 10.48550/ARXIV.2210.12084.

ALIANNEJADI M., KISELEVA J., CHUKLIN A., DALTON J. & BURTSEV M. (2021). Building and evaluating open-domain dialogue corpora with clarifying questions. In *EMNLP*.

ALIANNEJADI M., ZAMANI H., CRESTANI F. & CROFT W. B. (2019). Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, p. 475–484, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3331184.3331265.

ANAND A., CAVEDON L., HAGEN M., JOHO H., SANDERSON M. & STEIN B. (2020). Conversational search - a report from dagstuhl seminar 19461. *CoRR*, **abs/2005.08658**.

ASRI L. E., HE J. & SULEMAN K. (2016). A sequence-to-sequence model for user simulation in spoken dialogue systems. DOI : 10.48550/ARXIV.1607.00070.

BANERJEE S. & LAVIE A. (2005). METEOR : An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, p. 65–72, Ann Arbor, Michigan : Association for Computational Linguistics.

BELKIN N. J. & CROFT W. B. (1992). Information filtering and information retrieval : Two sides of the same coin ? *Commun. ACM*, **35**(12), 29–38. DOI : 10.1145/138859.138861.

BENAMARA F., HATOUT N., MULLER P. & OZDOWSKA S., Éds. (2007). *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, Toulouse. ATALA, IRIT.

BENDER E. M., GEBRU T., MCMILLAN-MAJOR A. & SHMITCHELL S. (2021). On the dangers of stochastic parrots : Can language models be too big ? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, p. 610–623, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3442188.3445922.

BI K., AI Q. & CROFT W. B. (2021). Asking clarifying questions based on negative feedback in conversational search. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '21, p. 157–166, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3471158.3472232.

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, Vancouver, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/P17-1171.

CHOI E., HE H., IYYER M., YATSKAR M., YIH W.-T., CHOI Y., LIANG P. & ZETTLEMOYER L. (2018). QuAC : Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 2174–2184, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/D18-1241.

CHU-CARROLL J. & BROWN M. K. (1997). Tracking initiative in collaborative dialogue interactions. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, p. 262–270, Madrid, Spain : Association for Computational Linguistics. DOI : 10.3115/976909.979651.

CLARKE C. L. A., CRASWELL N. & SOBOROFF I. (2009). Overview of the TREC 2009 web track. In E. M. VOORHEES & L. P. BUCKLAND, Éds., *Proceedings of The Eighteenth Text REtrieval Conference, TREC 2009, Gaithersburg, Maryland, USA, November 17-20, 2009*, volume 500-278 de *NIST Special Publication* : National Institute of Standards and Technology (NIST).

DALTON J., FISCHER S., OWOICHO P., RADLINSKI F., ROSSETTO F., TRIPPAS J. R. & ZAMANI H. (2022). Conversational information seeking : Theory and application. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, p. 3455–3458, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3477495.3532678.

DALTON J., XIONG C. & CALLAN J. (2020a). Trec cast 2019 : The conversational assistance track overview. DOI : 10.48550/ARXIV.2003.13624.

DALTON J., XIONG C., KUMAR V. & CALLAN J. (2020b). *CAsT-19 : A Dataset for Conversational Information Seeking*, In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 1985–1988. Association for Computing Machinery : New York, NY, USA.

DIAS G., Éd. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

DOU Z., JIANG Z., HU S., WEN J.-R. & SONG R. (2016). Automatically mining facets for queries from their search results. *IEEE Trans. on Knowl. and Data Eng.*, **28**(2), 385–397. DOI : 10.1109/TKDE.2015.2475735.

ELGOHARY A., PESKOV D. & BOYD-GRABER J. (2019). Can you unpack that ? learning to rewrite questions-in-context. 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. 5918–5924, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1605.

ERBACHER P., DENOYER L. & SOULIER L. (2022). : sigir. DOI : 10.48550/ARXIV.2205.15918.

FAN A., JERNITE Y., PEREZ E., GRANGIER D., WESTON J. & AULI M. (2019). ELI5 : Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 3558–3567, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1346.

FU Z., XIAN Y., ZHU Y., ZHANG Y. & DE MELO G. (2020). Cookie : A dataset for conversational recommendation over knowledge graphs in e-commerce. DOI : 10.48550/ARXIV.2008.09237.

GAO C., LI S., LEI W., CHEN J., LI B., JIANG P., HE X., MAO J. & CHUA T.-S. (2022). Kuairec : A fully-observed dataset and insights for evaluating recommender systems. In *Proceedings of the 31st ACM International Conference on Information amp ; Knowledge Management*, CIKM '22, p. 540–550, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3511808.3557220.

GLAESE A., MCALEESE N., TRĘBACZ M., ASLANIDES J., FIROIU V., EWALDS T., RAUH M., WEIDINGER L., CHADWICK M., THACKER P., CAMPBELL-GILLINGHAM L., UESATO J., HUANG P.-S., COMANESCU R., YANG F., SEE A., DATHATHRI S., GREIG R., CHEN C., FRITZ D., ELIAS J. S., GREEN R., MOKRÁ S., FERNANDO N., WU B., FOLEY R., YOUNG S., GABRIEL I., ISAAC W., MELLOR J., HASSABIS D., KAVUKCUOGLU K., HENDRICKS L. A. & IRVING G. (2022). Improving alignment of dialogue agents via targeted human judgements. DOI : 10.48550/ARXIV.2209.14375.

JIA M., LIU R., WANG P., SONG Y., XI Z., LI H., SHEN X., CHEN M., PANG J. & HE X. (2022). E-ConvRec : A large-scale conversational recommendation dataset for E-commerce customer service. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 5787–5796, Marseille, France : European Language Resources Association.

KANG D., BALAKRISHNAN A., SHAH P., CROOK P., BOUREAU Y.-L. & WESTON J. (2019a). Recommendation as a communication game : Self-supervised bot-play for goal-oriented dialogue. DOI : 10.48550/ARXIV.1909.03922.

KANG D., BALAKRISHNAN A., SHAH P., CROOK P., BOUREAU Y.-L. & WESTON J. (2019b). Recommendation as a communication game : Self-supervised bot-play for goal-oriented dialogue. 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. 1951–1961, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1203.

KEYVAN K. & HUANG J. X. (2022). How to approach ambiguous queries in conversational search : A survey of techniques, approaches, tools, and challenges. *ACM Comput. Surv.*, **55**(6). DOI : 10.1145/3534965.

KONG W. & ALLAN J. (2013). Extracting query facets from search results. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, p. 93–102, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/2484028.2484097.

KREYSSIG F., CASANUEVA I., BUDZIANOWSKI P. & GAŠIĆ M. (2018). Neural user simulation for corpus-based policy optimisation of spoken dialogue systems. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, p. 60–69, Melbourne, Australia : Association for Computational Linguistics. DOI : 10.18653/v1/W18-5007.

LAIGNELET M. & RIOULT F. (2009). Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs. In A. NAZARENKO & T. POIBEAU, Éds., *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis : ATALA LIPN.

LANGLAIS P. & PATRY A. (2007). Enrichissement d'un lexique bilingue par analogie. In (Benamara *et al.*, 2007), p. 101–110.

LI R., KAHOU S., SCHULZ H., MICHALSKI V., CHARLIN L. & PAL C. (2018a). Towards deep conversational recommendations. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, p. 9748–9758, Red Hook, NY, USA : Curran Associates Inc.

LI R., KAHOU S. E., SCHULZ H., MICHALSKI V., CHARLIN L. & PAL C. (2018b). Towards deep conversational recommendations. In *Advances in Neural Information Processing Systems 31 (NIPS 2018)*.

LI X., CHEN Y.-N., LI L., GAO J. & CELIKYILMAZ A. (2017). End-to-end task-completion neural dialogue systems. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 733–743, Taipei, Taiwan : Asian Federation of Natural Language Processing.

LIN J., MA X., LIN S.-C., YANG J.-H., PRADEEP R. & NOGUEIRA R. (2021). Pyserini : A Python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, p. 2356–2362.

LIU Z., WANG H., NIU Z.-Y., WU H. & CHE W. (2021). DuRecDial 2.0 : A bilingual parallel corpus for conversational recommendation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 4335–4347, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.356.

MOON S., SHAH P., KUMAR A. & SUBBA R. (2019). OpenDialKG : Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 845–854, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1081.

NAKANO R., HILTON J., BALAJI S., WU J., OUYANG L., KIM C., HESSE C., JAIN S., KOSARAJU V., SAUNDERS W., JIANG X., COBBE K., ELOUNDOU T., KRUEGER G., BUTTON K., KNIGHT M., CHESS B. & SCHULMAN J. (2021). Webgpt : Browser-assisted question-answering with human feedback. DOI : 10.48550/ARXIV.2112.09332.

NGUYEN T., ROSENBERG M., SONG X., GAO J., TIWARY S., MAJUMDER R. & DENG L. (2016). MS MARCO : A human generated machine reading comprehension dataset. *CoRR*, **abs/1611.09268**.

OUYANG L., WU J., JIANG X., ALMEIDA D., WAINWRIGHT C. L., MISHKIN P., ZHANG C., AGARWAL S., SLAMA K., RAY A., SCHULMAN J., HILTON J., KELTON F., MILLER L., SIMENS M., ASKELL A., WELINDER P., CHRISTIANO P., LEIKE J. & LOWE R. (2022). Training language models to follow instructions with human feedback. DOI : 10.48550/ARXIV.2203.02155.

OVER P. (2001). The trec interactive track : an annotated bibliography. *Information Processing Management*, **37**(3), 369–381. Interactivity at the Text Retrieval Conference (TREC), DOI : https ://doi.org/10.1016/S0306-4573(00)00053-4.

PENG B., LI X., GAO J., LIU J. & WONG K.-F. (2018). Deep Dyna-Q : Integrating planning for task-completion dialogue policy learning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 2182–2192, Melbourne, Australia : Association for Computational Linguistics. DOI : 10.18653/v1/P18-1203.

PRADEEP R., NOGUEIRA R. & LIN J. (2021). The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models. *CoRR*, **abs/2101.05667**.

QU C., YANG L., CHEN C., QIU M., CROFT W. B. & IYYER M. (2020). Open-retrieval conversational question answering. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, p. 539–548, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3397271.3401110.

QU Y., DING Y., LIU J., LIU K., REN R., ZHAO W. X., DONG D., WU H. & WANG H. (2021). RocketQA : An optimized training approach to dense passage retrieval for open-domain question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 5835–5847, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.naacl-main.466.

RADLINSKI F. & CRASWELL N. (2017). A theoretical framework for conversational search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, CHIIR '17, p. 117–126, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3020165.3020183.

RAFFEL C., SHAZEER N., ROBERTS A., LEE K., NARANG S., MATENA M., ZHOU Y., LI W. & LIU P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, **21**(140), 1–67.

RAJPURKAR P., JIA R. & LIANG P. (2018). Know what you don't know : Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 784–789, Melbourne, Australia : Association for Computational Linguistics. DOI : 10.18653/v1/P18-2124.

REDDY S., CHEN D. & MANNING C. D. (2019). CoQA : A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, **7**, 249–266. DOI : 10.1162/tacl_a_00266.

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.

SALLE A., MALMASI S., ROKHLENKO O. & AGICHTEIN E. (2021). Studying the effectiveness of conversational search refinement through user simulation. In D. HIEMSTRA, M.-F. MOENS, J. MOTHE, R. PEREGO, M. POTTHAST & F. SEBASTIANI, Éds., *Advances in Information Retrieval*, p. 587–602, Cham : Springer International Publishing.

SCHATZMANN J., THOMSON B., WEILHAMMER K., YE H. & YOUNG S. (2007). Agenda-based user simulation for bootstrapping a POMDP dialogue system. In *Human Language Technologies 2007 : The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, p. 149–152, Rochester, New York : Association for Computational Linguistics.

SEKULIĆ I., ALIANNEJADI M. & CRESTANI F. (2021). Towards facet-driven generation of clarifying questions for conversational search. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '21, p. 167–175, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3471158.3472257.

SEKULIĆ I., ALIANNEJADI M. & CRESTANI F. (2022). Exploiting document-based features for clarification in conversational search. In M. HAGEN, S. VERBERNE, C. MACDONALD, C. SEIFERT, K. BALOG, K. NØRVÅG & V. SETTY, Éds., *Advances in Information Retrieval*, p. 413–427, Cham : Springer International Publishing.

SERETAN V. & WEHRLI E. (2007). Collocation translation based on sentence alignment and parsing. In (Benamara *et al.*, 2007), p. 401–410.

SHAH C. & BENDER E. M. (2022). Situating search. In *ACM SIGIR Conference on Human Information Interaction and Retrieval*, CHIIR '22, p. 221–232, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3498366.3505816.

SHANNON C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, **27**, 379–423.

SHARMA P. & LI Y. (2019). Self-supervised contextual keyword and keyphrase retrieval with self-labelling. DOI : 10.20944/preprints201908.0073.v1.

SHAZEER N. & STERN M. (2018). Adafactor : Adaptive learning rates with sublinear memory cost. In J. DY & A. KRAUSE, Éds., *Proceedings of the 35th International Conference on Machine Learning*, volume 80 de *Proceedings of Machine Learning Research*, p. 4596–4604 : PMLR.

SHUSTER K., KOMEILI M., ADOLPHS L., ROLLER S., SZLAM A. & WESTON J. (2022). Language models that seek for knowledge : Modular search amp ; generation for dialogue and prompt completion. DOI : 10.48550/ARXIV.2203.13224.

THAKUR N., REIMERS N., RÜCKLÉ A., SRIVASTAVA A. & GUREVYCH I. (2021). BEIR : A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

THOPPILAN R., DE FREITAS D., HALL J., SHAZEER N., KULSHRESHTHA A., CHENG H.-T., JIN A., BOS T., BAKER L., DU Y., LI Y., LEE H., ZHENG H. S., GHAFOURI A., MENEGALI M., HUANG Y., KRIKUN M., LEPIKHIN D., QIN J., CHEN D., XU Y., CHEN Z., ROBERTS A., BOSMA M., ZHAO V., ZHOU Y., CHANG C.-C., KRIVOKON I., RUSCH W., PICKETT M., SRINIVASAN P., MAN L., MEIER-HELLSTERN K., MORRIS M. R., DOSHI T., SANTOS R. D., DUKE T., SORAKER J., ZEVENBERGEN B., PRABHAKARAN V., DIAZ M., HUTCHINSON B., OLSON K., MOLINA A., HOFFMAN-JOHN E., LEE J., AROYO L., RAJAKUMAR R., BUTRYNA A., LAMM M., KUZMINA V., FENTON J., COHEN A., BERNSTEIN R., KURZWEIL R., AGUERA-ARCAS B., CUI C., CROAK M., CHI E. & LE Q. (2022). Lamda : Language models for dialog applications. DOI : 10.48550/ARXIV.2201.08239.

TRIPPAS J. R., SPINA D., THOMAS P., SANDERSON M., JOHO H. & CAVEDON L. (2020). Towards a model for spoken conversational search. *Information Processing Management*, **57**(2), 102162. DOI : https ://doi.org/10.1016/j.ipm.2019.102162.

WILLIAMS R. J. & ZIPSER D. (1989). A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, **1**(2), 270–280. DOI : 10.1162/neco.1989.1.2.270.

WOLF T., DEBUT L., SANH V., CHAUMOND J., DELANGUE C., MOI A., CISTAC P., RAULT T., LOUF R., FUNTOWICZ M., DAVISON J., SHLEIFER S., VON PLATEN P., MA C., JERNITE Y., PLU J., XU C., SCAO T. L., GUGGER S., DRAME M., LHOEST Q. & RUSH A. M. (2019). Huggingface's transformers : State-of-the-art natural language processing. DOI : 10.48550/ARXIV.1910.03771.

WU F., QIAO Y., CHEN J.-H., WU C., QI T., LIAN J., LIU D., XIE X., GAO J., WU W. & ZHOU M. (2020). MIND : A large-scale dataset for news recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 3597–3606, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.331.

ZAMANI H., DUMAIS S., CRASWELL N., BENNETT P. & LUECK G. (2020a). *Generating Clarifying Questions for Information Retrieval*, In *Proceedings of The Web Conference 2020*, p. 418–428. Association for Computing Machinery : New York, NY, USA.

ZAMANI H., MITRA B., CHEN E., LUECK G., DIAZ F., BENNETT P. N., CRASWELL N. & DUMAIS S. T. (2020b). Analyzing and learning from user interactions for search clarification. *CoRR*, **abs/2006.00166**.

ZAMANI H., TRIPPAS J. R., DALTON J. & RADLINSKI F. (2022). Conversational information seeking. *CoRR*, **abs/2201.08808**.

ZHOU K., ZHOU Y., ZHAO W. X., WANG X. & WEN J.-R. (2020a). Towards topic-guided conversational recommender system. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain, December 8-11, 2020*.

ZHOU K., ZHOU Y., ZHAO W. X., WANG X. & WEN J.-R. (2020b). Towards topic-guided conversational recommender system. In *Proceedings of the 28th International Conference on Computational Linguistics*, p. 4128–4139, Barcelona, Spain (Online) : International Committee on Computational Linguistics. DOI : 10.18653/v1/2020.coling-main.365.