CoSPLADE : Adaptation d'un Modèle Neuronal Basé sur des Représentations Parcimonieuses pour la Recherche d'Information Conversationnelle

Nam Le Hai, Thomas Gerald, Thibault Formal, Jian-Yun Nie, Benjamin Piwowarksi, Laure Soulier

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

CoSPLADE : Adaptation d'un Modèle Neuronal Basé sur des Représentations Parcimonieuses pour la Recherche d'Information Conversationnelle

Nam Le Hai[1], Thomas Gerald[2], Thibault Formal[1,3], Jian-Yun Nie[4], Benjamin Piwowarksi[1], Laure Soulier[1,2]
(1) Sorbonne Université, CNRS, ISIR, F-75005 Paris, France
(2) Université Paris-Saclay, CNRS, SATT Paris Saclay, LISN, 91405, Orsay, France
(3) Naver Labs Europe, Meylan, France
(4) University of Montreal, Montreal, Canada
(5) Université Paris-Saclay, CNRS, LISN, 91405, Orsay, France
first.last @{sorbonne-universite.fr,lisn.fr,naverlabs.com},
nie@iro.umontreal.ca

RÉSUMÉ

La recherche conversationnelle est une tâche qui vise à retrouver des documents à partir de la question courante de l'utilisateur ainsi que l'historique complet de la conversation. La plupart des méthodes antérieures sont basées sur une approche multi-étapes reposant sur une reformulation de la question. Cette étape de reformulation est critique, car elle peut conduire à un classement sous-optimal des documents. D'autres approches ont essayé d'ordonner directement les documents, mais s'appuient pour la plupart sur un jeu de données contenant des pseudo-labels. Dans ce travail, nous proposons une technique d'apprentissage à la fois "légère" et innovante pour un modèle contextualisé d'ordonnancement basé sur SPLADE. En s'appuyant sur les représentations parcimonieuses de SPLADE, nous montrons que notre modèle, lorsqu'il est combiné avec le modèle de ré-ordonnancement T5Mono, obtient des résultats qui sont compétitifs avec ceux obtenus par les participants des campagnes d'évaluation TREC CAsT 2020 et 2021. Le code source de notre papier ECIR (Hai *et al.*, 2023) est disponible sur https://github.com/anonymous.

ABSTRACT

CoSPLADE : Contextualizing SPLADE for Conversational Information Retrieval.

Conversational search is a difficult task that aims at retrieving documents based not only on the current user query but also on the full conversation history. Most of the previous methods are built on a multi-stage ranking approach relying on query reformulation, a critical intermediate step that might lead to a sub-optimal retrieval. Other approaches have tried to use first-stage neural rankers, but are either zero-shot or rely on learning-to-rank based on a dataset with pseudo-labels. In this work, we propose an innovative lightweight learning technique to train a first-stage ranker based on SPLADE. By relying on SPLADE sparse representations, we show that, when combined with a second-stage ranker based on T5Mono, the results are competitive on the TREC CAsT 2020 and 2021 tracks. The source code of our ECIR paper (Hai *et al.*, 2023) is available at https://github.com/anonymous.

MOTS-CLÉS : Recherche d'Information, Recherche Conversationnelle, Ordonnancement Préalable.

KEYWORDS: Information Retrieval, Conversational Search, First-Stage Ranking.

1 Introduction

Avec le développement des assistants conversationnels comme Siri, Alexa ou Cortana, la Recherche d'Information (RI) Conversationnelle est devenue un domaine de recherche important (Culpepper *et al.*, 2018; Dalton *et al.*, 2019). Une recherche est effectuée au cours d'une session, à la manière d'une conversation naturelle : le besoin d'information de l'utilisateur est exprimé par une séquence d'interactions (questions ou questions puis réponses), introduisant ainsi des interdépendances complexes entre les différents "tours". Les modèles neuronaux de RI se sont avérés très performants pour cette tâche (Dalton *et al.*, 2020, 2021). Ils différent des travaux antérieurs car ils reposent sur une étape d'expansion de question basée sur l'historique (Zamani *et al.*, 2022), qui prend en compte toutes les questions passées et leurs réponses associées. Ce modèle de reformulation est généralement appris sur le jeu de données CANARD (Elgohary *et al.*, 2019), qui se compose d'une série de questions et de leurs réponses associées, ainsi que d'une question désambiguïsée – appelée *question oracle*. Cependant, s'appuyer sur une étape de reformulation est coûteux en calcul et peut s'avérer sous-optimal, comme souligné dans (Krasakis *et al.*, 2022; Lin *et al.*, 2021). (Krasakis *et al.*, 2022) utilisent ColBERT (Khattab & Zaharia, 2020) en inférence sans aucun entraînement du modèle ("zéro-shot"), en remplaçant la question par la séquence de questions, ce qui est clairement sous-optimal. (Lin *et al.*, 2021) proposent d'apprendre une représentation dense *contextualisée* de l'historique des questions, en optimisant une fonction d'ordonnancement sur un jeu de données composé de labels simulés. Le processus d'apprentissage est complexe (e.g., labels non fiables), et long.

Nous présentons ici notre participation à la campagne d'évaluation TREC CAsT (Dalton *et al.*, 2020, 2021), étendue dans une papier long à ECIR (Hai *et al.*, 2023). Nous proposons une approche faisant le lien entre l'ordonnancement préalable ("first-stage") et l'étape intermédiaire de reformulation. Notre modèle est une extension du modèle SPLADE (Formal *et al.*, 2021, 2022) basé sur l'apprentissage de représentations parcimonieuses, permettant ainsi d'extraire des mots pertinents pour la reformulation. Le processus d'apprentissage est "léger", car nous nous concentrons sur les questions et n'utilisons aucun jugement de pertinence. Nous alignons la représentation de la question contextualisée par la conversation à celle de la question désambiguïsée définie par l'*oracle*. Nous utilisons ensuite un modèle de ré-ordonnancement neuronal en tirant parti des représentations parcimonieuses pour fournir un contexte conversationnel sous la forme de mots-clés sélectionnés par SPLADE.

2 Modèle Contextualisé pour la Recherche d'Information Conversationnelle

La campagne d'évaluation TREC CAsT se concentre sur des sessions de recherche conversationnelle contenant environ 10 tours d'échanges. Chaque étape correspond à une question et la réponse canonique qui lui est associée [1], qui sert ensuite de contexte pour les questions futures. Pour chaque tour $n \leq N$, où N est le dernier tour de la conversation, nous désignons par q_n la question correspondante et a_n sa réponse canonique. Le contexte d'une question q_n au tour n correspond à l'ensemble des questions et réponses précédentes et éventuellement les réponses associées. L'objectif principal du challenge TREC CAsT est de retrouver, pour chaque question q_n et son contexte – c'est-à-dire les tours de la conversation – les passages pertinents d dans une collection de passages $\mathcal{D}$. Notre approche se base sur deux étapes d'ordonnancement.

1. Sélectionnée manuellement comme étant la réponse la plus pertinente d'un système de base.

2.1 Ordonnancement Préalable

Le modèle SPLADE (Formal *et al.*, 2021, 2022) estime le score d'un document en utilisant le produit scalaire entre la représentation parcimonieuse d'un document ($\hat{d}$) et d'une question ($\hat{q}$) : $s(\hat{q}, \hat{d}) = \hat{q} \cdot \hat{d}$. De manière similaire à (Lin *et al.*, 2021), nous supposons que la représentation du document a été bien entraînée dans le modèle original, sur la tâche standard de RI *ad-hoc*. La représentation $\hat{d}$ du document d est donc obtenue en utilisant le modèle SPLADE pré-entraîné, i.e. $\hat{d} = SPLADE(\texttt{[CLS]}\ d;\ \theta_{SPLADE})$, où θ_{SPLADE} correspond aux paramètres SPLADE originaux (Formal *et al.*, 2022) [2]. Ces paramètres ne sont pas modifiés au cours de l'apprentissage.

La représentation de la question contextualisée au tour n, désignée par $\hat{q}_{n,k}$, est obtenue par un nouveau modèle dérivé du modèle pré-entraîné SPLADE, en intégrant le contexte de la conversation, à savoir les questions précédentes et les réponses précédentes de la façon suivante :

$$\hat{q}_{n,k} = \hat{q}_n^{queries} + \hat{q}_{n,k}^{answers} \quad (1)$$

$$\hat{q}_n^{queries} = SPLADE(\texttt{[CLS]}\ q_n\ \texttt{[SEP]}\ q_1\ \texttt{[SEP]}\ \ldots\ \texttt{[SEP]}\ q_{n-1}; \theta_{queries}) \quad (2)$$

$$\hat{q}_{n,k}^{answers} = \frac{1}{k} \sum_{i=n-k}^{n-1} SPLADE(q_n\ \texttt{[SEP]}\ a_i; \theta_{answers,k}) \quad (3)$$

où $\hat{q}_n^{queries}$ encode la question actuelle dans le contexte de toutes les questions précédentes, et $\hat{q}_{n,k}^{answers}$ encode la question actuelle dans le contexte de k. Nous utilisons deux versions de SPLADE paramétrées par $\theta_{queries}$ pour l'historique complet des questions et $\theta_{answers,k}$ pour les réponses.

Entraînement. Nous proposons un entraînement basé sur deux fonctions de coût ayant pour objectif de rapprocher la représentation de la question estimée avec celle de la question oracle ainsi qu'avec le contexte de la conversation. La représentation $\hat{q}_n^*$ de la question oracle est obtenue en utilisant le modèle SPLADE original : $\hat{q}_n^* = SPLADE(q_n^*; \theta_{SPLADE})$.

La première composante de notre fonction de coût est basée sur une erreur des moindres carrés (MSE) qui compare la représentation de la question estimée avec la question oracle : $Loss_{MSE}(\hat{q}_{n,k}, \hat{q}_n^*) = MSE(\hat{q}_{n,k}, \hat{q}_n^*)$. Nous avons de plus ajouté une fonction de coût MSE asymétrique, conçue pour encourager l'expansion des termes à partir des réponses passées, ainsi qu'éviter d'introduire du bruit en restreignant les termes à ceux présents dans la question oracle q_n^* :

$$Loss_{asym}(\hat{q}_{n,k}^{answers}, \hat{q}_n^*) = \left(\max(\hat{q}_n^* - \hat{q}_{n,k}^{answers}, 0)\right)^2 \quad (4)$$

2.2 Ré-Ordonnancement

Nous effectuons le ré-ordonnancement en utilisant une approche T5Mono (Nogueira *et al.*, 2020), où nous enrichissons la question brute q_n avec des mots-clés identifiés par les représentations obtenues à l'issue de la première étape. La question enrichie q_n^+ pour le tour de conversation n est la suivante :

$$q_n^+ = q_n . Contexte : q_1\ q_2\ \ldots\ q_{n-1} . Mots - cles : w_1, w_2, ..., w_K \quad (5)$$

où les w_i sont les mots les plus importants du top-K que nous sélectionnons en exploitant le modèle d'ordonnancement préalable.

2. Qui peuvent être obtenus à partir de HuggingFace (Wolf *et al.*, 2020) : https://huggingface.co/naver/splade-cocondenser-ensembledistil

TREC CAsT 2020	Recall@1000	MAP@1000	MRR	nDCG@1000	nDCG@5	nDCG@3
TREC Participant (best)	63.3	30.2	59.3	52.6	-	45.8
TREC Participant (median)	52.1	15.1	42.2	36.4	-	30.4
TREC Participant (low)	27.9	1.0	5.9	11.1	-	2.2
CoSPLADE	82.4±2.0	26.9±1.5	58.1±2.9	54.2±1.8	41.2±2.4	44.0±2.7
TREC CAsT 2021	Recall@500	MAP@500	MRR	nDCG@500	nDCG@5	nDCG@3
TREC Participants 1 (best)	85.0	37.6	67.9	63.6	-	52.6
TREC Participants 2 (median)	36.4	17.6	53.4	33.6	-	37.7
TREC Participants 3 (low)	58.9	7.6	27.0	31.4	-	15.4
CoSPLADE	84.9±1.7	35.5±1.8	69.8±3	62.2±1.9	51.99±2.6	54.4±2.9

TABLE 1 – TREC CAsT 2020 and 2021 performances regarding participants

3 Évaluation et Résultats

Nous avons conçu le protocole d'évaluation de manière à satisfaire deux objectifs d'évaluation : *i)* Évaluer séparément l'efficacité des composantes d'ordonnancement des deux étapes de CoSPLADE ; *ii)* Comparer CoSPLADE avec les modèles des participants à TREC CAsT 2020 et 2021.

Pour entraîner notre modèle, nous avons utilisé le corpus CANARD, un ensemble de données conversationnelles axé sur la réécriture de questions basée sur le contexte. Plus précisément, le jeu de données CANARD est une liste d'historiques de conversations, chacune étant composée d'une série de questions, de réponses courtes (écrites par des humains) et de questions reformulées (contextualisées). Les ensembles d'entraînement, de développement et de test comprennent respectivement $31,538$, $3,418$ et $5,571$ questions contextuelles et reformulées.

Pour évaluer notre modèle, nous avons utilisé les ensembles de données TREC CAsT 2020 et 2021 qui comprennent respectivement 25 et 26 besoins en information ("topics") et une collection de documents composée de l'ensemble de données MS MARCO, d'une mise à jour de Wikipedia à partir du benchmark KILT (Petroni *et al.*, 2020) et de la collection Washington Post V4. Pour chaque besoin d'information, une conversation est disponible, alternant questions et réponses (passages sélectionnés manuellement dans la collection, i.e. réponses canoniques). Pour chaque question (216 et 239 au total), le jeu de données fournit sa forme réécrite manuellement ainsi qu'un ensemble d'environ 20 documents pertinents. Nous utilisons le premier pour définir une borne supérieure de comparaison.

L'analyse des différentes variantes de notre modèle ainsi que la comparaison avec des modèles de l'état de l'art et des participants TREC met en évidence les conclusions suivantes : 1) L'exploitation de l'historique des questions et des réponses permet de mieux contextualiser la question en cours. 2) Les réponses plus détaillées sont plus performantes. 3) Le coût asymétrique est bénéfique. 4) Notre modèle est capable d'obtenir des résultats comparables aux meilleurs modèles proposés par les participants TREC avec un modèle bien plus simple à entraîner et utilisant peu d'heuristiques.

4 Conclusion

Dans cet article, nous avons montré comment un modèle neuronal de RI basé sur des représentations parcimonieuses, à savoir SPLADE, pouvait être utilisé avec un processus d'apprentissage "léger" pour la RI conversationnelle. Nous avons obtenu des résultats comparables à ceux des systèmes les plus performants lors de la campagne d'évaluation TREC CAsT. Nous envisageons également d'évaluer notre approche sur d'autres jeux de données de QA conversationnelle, tels que CoQA (Reddy *et al.*, 2019), OR-ConvQA (Qu *et al.*, 2020), ou ConvMix (Christmann *et al.*, 2022).

5 Remerciements

Ce travail est financé par l'ANR JCJC SESAMS (ANR-18- CE23-0001) et l'ANR COST (ANR-18-CE23-0016).

Références

CHRISTMANN P., ROY R. S. & WEIKUM G. (2022). Conversational question answering on heterogeneous sources. In E. AMIGÓ, P. CASTELLS, J. GONZALO, B. CARTERETTE, J. S. CULPEPPER & G. KAZAI, Éds., *SIGIR '22 : The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, p. 144–154 : ACM. DOI : 10.1145/3477495.3531815.

CULPEPPER J. S., DIAZ F. & SMUCKER M. D. (2018). Research frontiers in information retrieval : Report from the third strategic workshop on information retrieval in lorne (SWIRL 2018). *SIGIR Forum*, **52**(1), 34–90. DOI : 10.1145/3274784.3274788.

DALTON J., XIONG C. & CALLAN J. (2019). TREC CAsT 2019 : The conversational assistance track overview. *arXiv*.

DALTON J., XIONG C. & CALLAN J. (2020). CAsT 2020 : The conversational assistance track overview. *arXiv*, p.1̃0.

DALTON J., XIONG C. & CALLAN J. (2021). TREC CAsT 2021 : The Conversational Assistance Track Overview. *arXiv*, p.7̃.

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.

FORMAL T., LASSANCE C., PIWOWARSKI B. & CLINCHANT S. (2022). From Distillation to Hard Negative Sampling : Making Sparse Neural IR Models More Effective. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, p. 2353–2359, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3477495.3531857.

FORMAL T., PIWOWARSKI B. & CLINCHANT S. (2021). SPLADE : Sparse Lexical and Expansion Model for First Stage Ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, p. 2288–2292, New York, NY, USA : Association for Computing Machinery. DOI : 10/gm2tf2.

HAI N. L., GERALD T., FORMAL T., NIE J., PIWOWARSKI B. & SOULIER L. (2023). Cosplade : Contextualizing SPLADE for conversational information retrieval. In *Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Proceedings, Part I*, p. 537–552. DOI : 10.1007/978-3-031-28244-7_34.

KHATTAB O. & ZAHARIA M. (2020). ColBERT : Efficient and effective passage search via contextualized late interaction over BERT. *arXiv*.

KRASAKIS A. M., YATES A. & KANOULAS E. (2022). Zero-shot Query Contextualization for Conversational Search. In *Proceedings of the 45th International ACM SIGIR Conference on*

Research and Development in Information Retrieval, SIGIR '22, p. 1880–1884, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3477495.3531769.

LIN S.-C., YANG J.-H. & LIN J. (2021). Contextualized query embeddings for conversational search. *arXiv*.

NOGUEIRA R., JIANG Z., PRADEEP R. & LIN J. (2020). Document ranking with a pretrained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics : EMNLP 2020*, p. 708–718 : Association for Computational Linguistics. DOI : 10.18653/v1/2020.findings-emnlp.63.

PETRONI F., PIKTUS A., FAN A., LEWIS P., YAZDANI M., DE CAO N., THORNE J., JERNITE Y., KARPUKHIN V., MAILLARD J. *et al.* (2020). Kilt : a benchmark for knowledge intensive language tasks. *arXiv preprint arXiv :2009.02252*.

QU C., YANG L., CHEN C., QIU M., CROFT W. B. & IYYER M. (2020). Open-retrieval conversational question answering. In J. X. HUANG, Y. CHANG, X. CHENG, J. KAMPS, V. MURDOCK, J. WEN & Y. LIU, Éds., *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, p. 539–548 : ACM. DOI : 10.1145/3397271.3401110.

REDDY S., CHEN D. & MANNING C. D. (2019). Coqa : A conversational question answering challenge. *Trans. Assoc. Comput. Linguistics*, **7**, 249–266. DOI : 10.1162/tacl_a_00266.

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. (2020). Transformers : State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, p. 38–45, Online : Association for Computational Linguistics.

ZAMANI H., TRIPPAS J. R., DALTON J. & RADLINSKI F. (2022). Conversational Information Seeking. arXiv :2201.08808 [cs], DOI : 10.48550/arXiv.2201.08808.