Adaptation de domaine pour la recherche dense par annotation automatique

Minghan Li, Eric Gaussier

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

Adaptation de domaine pour la recherche dense par annotation automatique

Minghan Li[1] Eric Gaussier[1]
(1) Univ. Grenoble Alpes, CNRS, LIG, Bâtiment IMAG - 700 avenue Centrale, 38000 Grenoble, France
minghan.li@univ-grenoble-alpes.fr, eric.gaussier@imag.fr

RÉSUMÉ

Bien que la recherche d'information neuronale ait connu des améliorations, les modèles de recherche dense ont une capacité de généralisation à de nouveaux domaines limitée, contrairement aux modèles basés sur l'interaction. Les approches d'apprentissage adversarial et de génération de requêtes n'ont pas résolu ce problème. Cet article propose une approche d'auto-supervision utilisant des étiquettes de pseudo-pertinence automatiquement générées pour le domaine cible. Le modèle T53B est utilisé pour réordonner une liste de documents fournie par BM25 afin d'obtenir une annotation des exemples positifs. L'extraction des exemples négatifs est effectuée en explorant différentes stratégies. Les expériences montrent que cette approche aide le modèle dense sur le domaine cible et améliore l'approche de génération de requêtes GPL.

ABSTRACT

Domain adaptation with pseudo-relevance labeling for dense retrieval.

Although neural information retrieval has witnessed great improvements, recent works showed that the generalization ability of dense retrieval models on target domains with different distributions is limited, which contrasts with the results obtained with interaction-based models. To address this issue, researchers have resorted to adversarial learning and query generation approaches ; both approaches nevertheless resulted in limited improvements. In this paper, we propose to use a self-supervision approach in which pseudo-relevance labels are automatically generated on the target domain. To do so, we use the interaction-based model T53B to re-rank the BM25 list on target domain for pseudo positive labeling. Since negative mining is vital, we carefully design it with investigating different negative mining strategies. Our experiments reveal that the proposed pseudo-relevance labeling approach helps the dense retrieval model on target domain and improves the state-of-the-art query generation approach GPL when they are fine-tuned on the generated data.

MOTS-CLÉS : Adaptation de domaine, Apprentissage auto-supervisé, Recherche d'information neuronale.

KEYWORDS: Dense Retrieval, Domain Adaptation, Self-Supervised Learning, Neural IR.

1 Introduction

La recherche d'information (RI) joue un rôle crucial dans notre vie quotidienne en raison de l'explosion des données. Les approches traditionnelles de RI telles que BM25 (40) calculent une similarité entre une requête et un document uniquement sur la base des termes communs aux deux. En tant que telles, elles ne peuvent pas gérer la correspondance sémantique entre différentes formes de surface.

La recherche d'informations neuronale, avec l'avènement des réseaux de neurones profonds, a considérablement amélioré les systèmes de RI grâce à des modèles capables de capturer la sémantique de chaque terme et de les comparer même si leur forme de surface diffère. Un modèle populaire à la fois en traitement du langage naturel (NLP) et en RI est BERT (4), qui est basé sur des transformateurs (47) et est pré-entraîné sur de grandes collections ; BERT peut être utilisé sur une variété de tâches secondaires par adaptation (en anglais *fine-tuning*).

Les modèles de RI neuronale peuvent être classés en deux catégories (11) : les approches basées sur l'interaction et les approches basées sur la représentation (également appelées recherche dense). Les modèles basés sur l'interaction ont montré des performances moyennes supérieures à celles des modèles de recherche dense ; les modèles de recherche dense sont toutefois plus rapides que les modèles basés sur l'interaction, car les représentations des documents peuvent être générées et stockées à l'avance, et sont préférables si l'on a besoin de déployer un modèle à grande échelle. Cela étant dit, des études récentes comme BEIR (44) ont montré que les modèles de recherche dense entraînés sur un domaine source généralisent moins bien que les modèles traditionnels tels que BM25 et les modèles basés sur l'interaction sur des ensembles de données hors distribution (OOD pour *Out Of Distribution*). Bien que l'entraînement sur des ensembles de données cibles avec des étiquettes de référence soit un processus standard, l'annotation requise peut être à la fois longue et coûteuse, de sorte que cette approche peut ne pas être applicable à de nombreuses utilisations réelles. Il est donc important de traiter le problème dans des scénarios OOD pour la recherche dense.

L'un des objectifs de l'adaptation de domaine (53; 50) est de permettre à un modèle entraîné sur un domaine appelé le domaine source de bien fonctionner sur un autre domaine appelé le domaine cible sans utiliser d'étiquettes humaines sur ce dernier. Récemment, différentes techniques d'adaptation de domaine pour la recherche dense ont été proposées. La généralisation de domaine basée sur la génération de données est l'une des approches (50) qui a été suivie dans (26) grâce à un modèle appelé QGen qui génère des requêtes pour le domaine cible en utilisant un générateur de requêtes entraîné sur le domaine source. Dans la même lignée, GPL (52) utilise des exemples négatifs difficiles et la distillation de connaissances et obtient des résultats de pointe sur un certain nombre d'ensembles de données BEIR. Cependant, les requêtes créées sont synthétiques et peuvent ne pas ressembler à de véritables requêtes cibles. Une autre approche populaire et largement utilisée est basée sur l'apprentissage adversarial (50). Très récemment, Xin et al. (54) ont proposé un modèle appelé MoDIR qui entraîne de manière adverse un encodeur de recherche dense pour apprendre des représentations invariantes aux domaines. Cependant, un tel objectif d'apprentissage peut produire un espace de plongement (en anglais *embedding*) de faible qualité et entraîner des performances instables (52; 14).

Dans cet article, nous proposons une approche dénommée DoDress (pour **Do**main generalization for **D**ense **re**trieval through **s**elf-**s**upervision) qui cherche d'abord à construire des annotations de pseudo-pertinence [1] sur le domaine cible en utilisant des modèles basés sur l'interaction uniquement entraînés sur le domaine source tels que T53B (34). La raison d'utiliser des modèles basés sur l'interaction dans ce contexte réside dans le fait que ces modèles ont montré un comportement relativement bon sur les ensembles de données OOD (44). Notez que le modèle T53B lourd n'est utilisé que pour produire des étiquettes de pseudo-relation avant l'entraînement du modèle de recherche dense afin que l'approche globale reste efficace pendant la recherche en ligne. Cette méthode élimine le besoin d'annotations humaines et permet au modèle d'utiliser de véritables requêtes et documents du domaine cible.

Une des difficultés dans l'annotation automartique est d'obtenir des exemples négatifs de qualité.

1. Nous utilisons ce terme pour rendre compte du fait que certaines de ces annotations sont erronées.

Nous étudions pour cela dans cet article différentes stratégies : l'échantillonnage négatif aléatoire global, les exemples négatifs "durs" de BM25 et les exemples négatifs "durs" de SimANS (16).

Notre contribution réside dans l'étude et la combinaison de différentes approches existantes pour l'annotation automatique dans un domaine cible qui conduit *n fine* à une méthode qui améliore l'état de l'art pour l'adaptation de domaine de modèles denses.

2 Travaux connexes

Wang *et al.* (50) présentent un article de synthèse sur la généralisation de domaine pour les domaines non vus. La généralisation ou l'adaptation de domaine peut être catégorisée en trois groupes : manipulation de données, apprentissage de représentation et stratégie d'apprentissage. Il existe deux types de techniques dans le premier groupe : l'augmentation de données (35; 45; 43; 48) qui est couramment utilisée dans les données d'images (par exemple, en modifiant la localisation, le texte des objets et en ajoutant du bruit aléatoire) et la génération de données (38; 36; 57) qui utilise certains modèles pour générer de nouvelles données pour entraîner un modèle. Le groupe d'apprentissage de représentation comprend l'apprentissage de représentation invariante de domaine (par exemple, l'apprentissage adversarial de domaine) (1; 7; 31) et les méthodes de désentrelacement de fonctionnalités (21; 32; 24). Le troisième groupe comporte plusieurs catégories, comme l'apprentissage en ensemble (28; 6), l'apprentissage méta (20; 5) et les approches basées sur l'apprentissage auto-supervisé (par exemple, la résolution de puzzles de type jigsaw) (3; 13).

Des stratégies similaires, telles que la généralisation de domaine ou l'apprentissage par transfert, sont avancées par les chercheurs pour la recherche d'informations. Une stratégie similaire à celle adoptée dans cette étude est décrite dans (30), qui effectue une évaluation systématique de la capacité de transfert des modèles de classement neuronaux basés sur BERT. Les auteurs utilisent également BM25 pour générer des étiquettes de pseudo-pertinence. Ils ne se concentrent cependant pas sur les modèles de recherche denses qui sont connus pour nécessiter des méthodes d'entraînement complexes et une grande quantité de données dans une situation distincte (8). De plus, l'utilisation uniquement de BM25 pour obtenir des étiquettes de pseudo-pertinence pourrait être une solution faible. Pour les modèles basés sur l'interaction (19) ou l'apprentissage d'embeddings de phrases (10; 51), certaines publications suggèrent des techniques auto-supervisées. Ces méthodes sont fréquemment utilisées pour la pré-formation, mais elles ne se concentrent pas explicitement sur la généralisation de domaine (52). Ma *et al.* (26) propose QGen, une approche de génération d'apprentissage zéro pour la première étape de la recherche dense de passages qui utilise la génération de questions synthétiques, permettant la construction de paires de pertinence question-passage arbitrairement grandes mais bruyantes qui sont spécifiques au domaine, dans le but de surmonter le défi que les modèles de recherche neuronaux ont besoin d'un grand ensemble d'entraînement supervisé pour surpasser les approches conventionnelles basées sur les termes. Les documents utilisés pour générer les requêtes sont considérés comme des passages positifs et les autres instances dans le lot sont considérées comme négatives. En parallèle, Liang *et al.* (23) examinent l'architecture de recherche de passages denses à deux tours. Étant donné que les données étiquetées peuvent être difficiles à obtenir et que les modèles de recherche neuronaux ont besoin d'une grande quantité de données pour être entraînés, ils suggèrent également d'utiliser des requêtes synthétiques produites par un grand modèle de séquence à séquence (seq2seq) pour l'adaptation de domaine non supervisée. Ces deux articles montrent l'efficacité de l'approche de génération de requêtes, qui est également utilisée dans le modèle GPL (52). GPL

s'appuie sur un encodeur-décodeur T5 pré-entraîné (37) pour générer des requêtes à partir de passages d'entrée. Les passages d'entrée sont considérés comme des passages positifs tandis que les passages similaires récupérés à l'aide d'un modèle de recherche dense existant sont constitués de passages négatifs (difficiles). La perte Margin-MSE (12) est utilisée comme distillation de connaissances pour enseigner au modèle de recherche dense à apprendre à partir d'un modèle basé sur l'interaction. Les résultats expérimentaux montrent une efficacité de pointe sur plusieurs collections BEIR (44).

Les chercheurs ont également exploré des stratégies alternatives pour l'adaptation de domaine des modèles de recherche dense. Xin *et al.* (54) a proposé une approche d'apprentissage de représentation invariante de domaine adversaire avec une méthode d'impulsion pour l'apprentissage de classifier de domaine qui distingue les domaines source et cible. L'encodeur de recherche dense est ensuite formé de manière adversaire pour apprendre des représentations invariantes de domaine. Une file d'attente à impulsion qui enregistre des embeddings de plusieurs lots précédents est utilisée afin de trouver un équilibre entre précision et efficacité (54). Cette approche est utilisée sur un modèle ANCE entraîné (55). Les résultats varient d'un ensemble de données à l'autre, avec parfois des améliorations importantes et parfois des gains ou pertes marginales. Karouzos *et al.* (14) a proposé UDALM pour l'adaptation de domaine pour la classification de sentiment à travers l'apprentissage multi-tâche. Il apprend simultanément l'objectif de la tâche de modélisation de langage masquée (MLM) sur le domaine cible et la tâche à partir des données étiquetées source. Cependant, cette stratégie n'a pas été conçue pour la recherche dense et, comme mentionné dans (52), elle ne fonctionne pas bien pour la recherche dense.

Dans cet article, nous proposons de faire l'adaptation de domaine pour la recherche dense par auto-supervision par étiquetage de pseudo-relevance. Nous appliquons le modèle d'interaction T53B de pointe et généralisable au domaine (34) pour l'étiquetage de pseudo-positif. Ce modèle peut produire des étiquettes de pseudo-relevance plus précises, où les documents classés en haut sont considérés comme pertinents pour une requête donnée. De plus, différentes stratégies d'échantillonnage négatives sont étudiées, en particulier avec les négatifs durs de SimANS (16) échantillonnés à partir de la liste de recherche des modèles DR actuels, pour améliorer l'efficacité du modèle, après la formation avec les données pseudo-étiquetées générées.

3 Contexte

Recherche dense La recherche dense (15; 54) vise à encoder à la fois les requêtes et les documents dans un espace de faible dimension à l'aide d'un encodeur g, généralement un modèle de type BERT. Le score de pertinence (RSV) d'une requête et d'un document est ensuite calculée à l'aide d'une fonction de similarité simple dans l'espace de faible dimension :

$$RSV(q,d)_{DR} = g(q) \cdot g(d) \quad (or\ RSV(q,d)_{DR} = cos(g(q), g(d)))\,, \tag{1}$$

où $g(q)$ (resp. $g(d)$) représente l'encodage de la requête (resp. du document). Cela permet une recherche rapide en utilisant par exemple la méthode proposée par Xiong *et al.* (55).

BM25 BM25 est un algorithme standard en RI basé sur la correspondance de termes. Le RSV d'un document par rapport à une requête est donné par :

$$RSV(q,d)_{BM25} = \sum_{w \in q \cap d} IDF(w) \cdot \frac{tf_w}{k_1 \cdot (1 - b + b \cdot \frac{l_d}{l_{avg}}) + tf_w}, \tag{2}$$

où $IDF(w)$ est la fréquence inverse des documents, l_d est la longueur du document d, l_{avg} est la longueur moyenne des documents dans l'ensemble de données, et k_1 et b sont deux hyperparamètres.

T53B T5 (37) est un modèle qui a montré son efficacité dans diverses tâches du traitement automatique des langues. Nogueira *et al.* (34) ont proposé d'utiliser T5 en tant que modèle basé sur l'interaction pour la recherche d'informations en se basant sur la représentation d'entrée suivante :

Query : [q] Document : [d] Relevant : true or false

où $[q]$ et $[d]$ sont remplacés par les textes de la requête et du document. Pendant l'entraînement, le modèle T5 apprend à générer le mot "true" lorsque le document est pertinent pour la requête, et le mot "false" lorsqu'il ne l'est pas. Le score de pertinence pour l'inférence est ensuite déterminé par la probabilité de produire "true" (34) :

$$RSV(q,d)_{T5} = softmax(Z_{true}) = \frac{e^{Z_{true}}}{e^{Z_{true}} + e^{Z_{false}}}, \tag{3}$$

où Z_{true} et Z_{false} sont les logits des tokens de sortie.

4 DoDress : annotation automatique de pertinence

Nous proposons simplement ici de considérer les k meilleurs documents, obtenus avec la combinaison BM25&T53B dans laquelle T53B ré-ordonne les documents fournis par BM25, comme pertinents. k est un hyperparamètre qui peut être ajusté en fonction de différentes informations, telles que le nombre de requêtes et de documents disponibles. Pour chaque paire (requête, document pertinent) obtenue, nous cherchons à extraire de la collection m documents non pertinents pour la requête. Ainsi, pour chaque requête, $k \times m$ triplets (requête, document pertinent, document non pertinent) sont consituées. Les blocs verts dans les Figures 1 et 2 représentent ces triplets qui constituent les données d'entraînement sur le domaine cible.

Une stratégie simple d'extraction de documents non pertinents consiste à un échantillonnage aléatoire global de la collection excluant les documents jugés pertinents. Toutefois, les modèles de recherche denses nécessitent des stratégies d'apprentissage complexes pour être performants (9) et l'approche précédente ne garantit pas que les documents non pertinents obtnues soient suffisamment informatifs. Un des défis clé pour la recherche dense est en effet de construire des instances négatives appropriées pour l'apprentissage (15). Une solution possible ici est d'utiliser des documents non pertinents proches des documents pertinents en termes de recherche obtenus par BM25. Nous échantillonons $k \times m$ documents parmi les documents de haut rang de BM25, en excluant bien sûr les k documents jugés pertinents après ré-ordonnancement par T53B, et les considérons comme non pertinents. Cette approche est illustrée dans la Figure 1 : les documents non pertinents sont échantillonnés au hasard

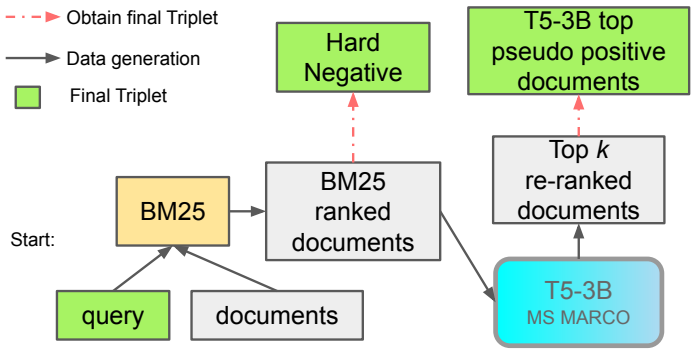

FIGURE 1 – Processus global d'annotation automatique avec échantillonnage négatif sur les résultats de BM25.

dans la liste de haut rang de BM25 alors que les documents pertinents correspondent aux k prem :iers documents obtenus après ré-ordonnancement par T53B.

L'approche précédente d'annotation est basée sur les instances les mieux classées et sur des exemples négatifs aléatoires globaux ou des exemples négatifs difficiles de BM25. Bien que les résultats soient globalement bons, des chercheurs ont récemment montré (16) que les stratégies d'échantillonnage négatif existantes souffrent du problème de faux négatifs ou d'informations non pertinentes, et ils montrent que les exemples négatifs classés autour des exemples positifs (par exemple, les scores BM25 ou les scores de recherche dense) sont généralement plus informatifs et moins susceptibles d'être des faux négatifs. Dans cet article, nous utilisons SimANS (16) comme illustré dans la figure 2. SimANS permet de sélectionner les documents non pertinents dans les classements des modèles denses D-BERT et GPL (voir Section **??**). À noter que les documents non pertinents classés autour de documents jugés pertinents ne sont pas forcément en tête de liste car les classements de D-BERT et GPL diffèrent de ceux de T53B. Toutefois, afin de ne pas sélectionner des documents trop mal classés, nous nous concentrons sur les Top500 documents fournis par D-BERT et GPL respectivement. Cela signifie que certains documents jugés pertinents sont susceptibles de ne plus être considérés s'ils n'appartiennet pas à ce Top500.

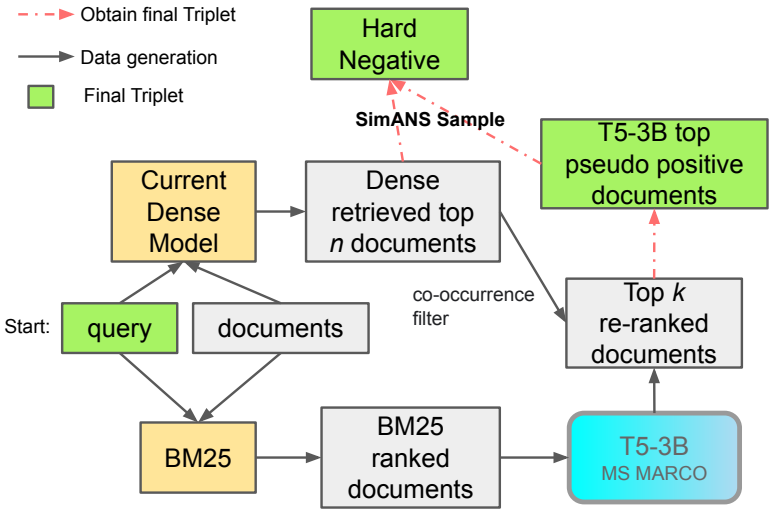

FIGURE 2 – Processus global d'annotation avec SimANS.

4.1 Combinaison avec GPL

Comme mentionné précédemment, les approches QGen et GPL s'appuient toutes deux sur un générateur de requêtes pour générer des pseudo-requêtes afin de construire un modèle de recherche dense. Nous proposons ici de construire un tel modèle sur les triplets de pseudo-pertinence décrits précédememnt et obtenus à partir des requêtes fournies par QGen ou GPL. Nous pensons qu'il est possible de tirer profit de cette formation supplémentaire sur la collection cible, car les pseudo-requêtes et les étiquettes de pseudo-pertinence reposent sur des sources d'informations différentes et sont complémentaires l'une de l'autre. Comme nous le verrons dans la section expérimentale, cette combinaison améliore effectivement l'approche de génération de pseudo-requêtes.

4.2 Fonction de coût

Dans cet article, nous nous appuyons sur la perte paire RankNet (2; 22) pour entraîner un modèle de recherche dense en utilisant les triplets générés ci-dessus, définis par :

$$\mathcal{L}(q, d^+, d^-; \Theta) = -\log(\sigma(S_{q,d^+} - S_{q,d^-})), \tag{4}$$

où q est une requête, (d^+, d^-) est une paire de documents d'entraînement (positif, négatif) pour q, σ est la fonction sigmoïde, Θ représente les paramètres du modèle de recherche dense, et $S_{q,d}$ est le score fourni par le modèle pour le document d par rapport à la requête q.

5 Expérimentation

5.1 Ensembles de données

Le jeu de données MS MARCO (33) est utilisé comme données de domaine source. Nous voulons expérimenter dans un scénario extrême où aucune requête de test ne peut être vue, même sans étiquettes humaines. Cela signifie que nous devons générer les données d'entraînement avec les requêtes d'entraînement qui ne sont pas dans l'ensemble de test. Pour ce faire, nous expérimentons sur 3 ensembles de données de domaine cible du benchmark BEIR (44) : FiQA, ensemble de questions-réponses sur la finance (27) qui contient 6000 requêtes d'entraînement, BioASQ, ensemble de questions-réponses dans le domaine biomédical (46) (suivant (52), les documents non pertinents sont éliminés au hasard pour ne conserver qu'un million de documents) qui contient 3243 requêtes d'entraînement provenant de la collection originale [2], et Robust04, ensemble documents d'actualités (49) qui contient 250 requêtes. Différents sujets et tâches sont couverts par ces ensembles. Pour Robust04, nous sélectionnons les 100 premières requêtes comme ensemble d'entraînement et de développement ; les 150 dernières requêtes sont utilisées comme ensemble de test.

2. http://participants-area.bioasq.org/Tasks/8b/trainingDataset/

5.2 Protocole expérimental

Notre implémentation est basée sur le cadre open-source Matchmaker [3] avec pooling moyen et évaluation dense [4] en utilisant la précision mixte automatique (29). Sur la collection cible, les triplets d'entraînement sont générés selon les approches décrites ci-dessus. Pour l'adaptation de domaine, nous utilisons deux modèles denses, D-BERT et GPL. D-BERT correspond au modèle DistilBERT (41) avec 6 couches. GPL est d'abord entraîné sur les pseudo-requêtes cibles qu'il génère et les documents associés avant d'être entraîné sur les triplets cibles. Le modèle T5 utilisé est la version 3B qui est entraînée sur l'ensemble de données de classement de passage MS MARCO [5]. Le *cross-encoder* MiniLM utilisé est la version *ms-marco-MiniLM-L-6-v2* [6] de *sentence transformers* (39).

Pour construire l'ensemble d'entraînement, nous sélectionnons le nombre k de documents principaux à considérer comme pertinents en fonction du nombre de requêtes (pour générer suffisamment de paires) et de documents (un grand nombre de documents permettant d'échantillonner plus de documents négatifs). Pour chaque document pertinent, nous sélectionnons m documents de la collection cible qui ne figurent pas dans la liste des k premiers documents de la requête selon les différentes stratégies d'échantillonnage négatif présentées précédemment. Ces documents sont considérés comme non pertinents. Le Tableau 1 affiche le nombre de requêtes, la valeur sélectionnée pour k (entre parenthèses) et le nombre m de documents non pertinents par document pertinent. Par exemple, pour BioASQ, le nombre de triplets dans l'ensemble d'entraînement est de $3193 \times 2 \times 15 = 95790$. À la fin, chaque ensemble de données a un nombre de triplets dans l'ensemble d'entraînement compris entre 50000 et 100000. Nous construisons également un ensemble de développement pour sélectionner les hyperparamètres des modèles sur chaque collection. Pour chaque requête de l'ensemble de développement, les 10 premiers documents sont considérés comme pertinents et 90 documents sélectionnés au hasard comme non pertinents. Ce choix est dicté par le fait que nous avons besoin d'un nombre suffisant de documents pertinents à des fins d'évaluation et que nous avons un nombre limité de requêtes pour l'ensemble de développement. Cependant, pour contrebalancer le risque de considérer comme pertinents des documents qui ne le sont pas en réalité, les deux premiers documents sont étiquetés "2" et les huit suivants comme "1". Les documents non pertinents sont étiquetés "0", ce qui conduit à des jugements de pertinence à 3 niveaux pour chaque ensemble de données. Le meilleur modèle est sauvegardé en fonction du score NDCG@10 sur l'ensemble de développement évalué toutes les 1 000 étapes.

Suivant (52), une longueur de séquence maximale de 350 et une similarité par produit scalaire sont utilisées. Pour tous les ensembles de données, nous utilisons une taille de lot de 8, ce qui signifie 8 paires positives-négatives, et un taux d'apprentissage de 2e-6 avec un optimiseur Adam pour 10 000 étapes d'entraînement. Un schéma LR cosinus (25) est également utilisé pour la décroissance du taux d'apprentissage.

Pour SimANS, les hyperparamètres a et b sont fixés à 0,5 et 0 respectivement pour toutes les expériences.

3. https://github.com/sebastian-hofstaetter/matchmaker
4. https://github.com/UKPLab/gpl/blob/main/gpl/toolkit/evaluation.py
5. https://huggingface.co/castorini/monot5-3b-msmarco
6. https://huggingface.co/cross-encoder/ms-marco-MiniLM-L-6-v2

TABLE 1 – Nombre de requêtes, de documents et de documents pertinents et non pertinents par requête pour chaque collection. Les nombres entre parenthèses correspondent aux requêtes utilisées pour l'entraînement.

dataset	#requêtes	#docs	k	m
FiQA	6000 (5960)	57K	1	10
BioASQ	3243 (3193)	1M	2	15
Robust04	100 (90)	528K	15	67

5.3 Modèles *baseline*

Conformément à (52), nous comparons les approches proposées avec des modèles sans apprentissage, avec des approches de pré-entraînement et avec les approches récentes les plus performantes pour l'adaptation de domaine. Les résultats des modèles basés sur l'interaction sont également présentés.

5.3.1 Modèles sans apprentissage

Les modèles de référence sans apprentissage comprennent BM25 basé sur Anserini (56) avec des paramètres par défaut qui obtient les 100 meilleurs documents pour chaque requête et ne nécessite pas d'être entraîné (ces listes de classement BM25 sont ensuite utilisées pour générer des données d'apprentissage de pseudo-pertinence dans cet article) et le modèle de recherche dense D-BERT entraîné uniquement sur la collection source, la fonction de coût MarginMSE et le modèle *ms-marco-MiniLM-L-6-v2* pris comme modèle enseignant (il est ensuite utilisé comme point de départ pour l'adaptation de domaine).

5.3.2 Modèles basés sur le pré-entraînement

Nous comparons avec SimCSE (10), ICT (19) et TSDAE (51). Ces modèles sont tous d'abord pré-entraînés de manière auto-supervisée sur l'ensemble de données cible, puis affinés sur MS MARCO.

5.3.3 Approches d'adaptation de domaine

Nous comparons ici quatre approches SOTA récentes : MoDIR (54), qui repose sur ANCE (55) et utilise un entraînement adversarial, UDALM (14), qui repose sur l'apprentissage multi-tâches, et QGen (26) et GPL (52), qui sont des approches basées sur la génération de requêtes.

De plus, nous utilisons les modèles basés sur l'interaction BM25+CE et BM25+T53B, qui peuvent être considérés comme des baselines solides en raison du bon comportement des modèles basés sur l'interaction dans les contextes OOD (44), mais qui restent néanmoins inefficaces lors de l'inférence. Ces modèles réordonnent la liste des 100 premiers résultats renvoyés par BM25, en utilisant respectivement les *corss-encoders ms-marco-MiniLM-L-6-v2* et T53B.

5.4 Résultats et analyse

Les tableaux 3, 4 et 5 affichent les résultats obtenus avec les différents modèles et approches. Les résultats rapportés pour BM25+CE, UDALM, MoDIR, SimCSE, ICT, TDSAE, QGen et TSDAE+GPL proviennent de (52). Comme nous testons le dataset Robust04 sur les 150 dernières requêtes, pour BM25+CE, GPL et TSDAE+GPL, nous chargeons les points de contrôle entraînés de D-BERT et de Wang *et al.* (52) [7], et les évaluons sur les 150 dernières requêtes. La notation "DoDress-BM25 (D-BERT)" (respectivement "DoDress-T53B (D-BERT)") correspond au modèle de recherche dense D-BERT pré-entraîné sur MS MARCO et affiné sur les données cibles en utilisant les documents pertinents obtenus par BM25 (respectivement en utilisant BM25+T5). La notation (GPL) signifie la même chose pour GPL, qui est d'abord entraîné sur les pseudo-requêtes cibles qu'il génère et les documents associés avant d'être entraîné sur les triplets cibles.

Nous analysons les résultats en répondant à trois questions de recherche.

RQ1 Les top positifs de BM25+T53B aident-ils à la généralisation de domaine pour les modèles de recherche dense ?

À partir du Tableau 3, on constate que DoDress-T53B (D-BERT) améliore D-BERT sur le jeu de données FiQA avec les trois stratégies d'échantillonnage négatives, et que DoDress-T53B (GPL) montre une tendance similaire par rapport à GPL. Sur le jeu de données Robust04, à partir du Tableau 4, nous observons des tendances similaires pour DoDress-T53B (D-BERT) et DoDress-T53B (GPL). Plus précisément, DoDress-T53B (GPL) avec les trois stratégies d'échantillonnage permet d'améliorer GPL et TSDAE + GPL. Avec la stratégie de recherche de négatifs SimANS, DoDress-T53B (D-BERT) montre une amélioration de 11,5% $((43,6 - 39,1) \div 39,1)$ par rapport à D-BERT, et DoDress-T53B (GPL) montre une amélioration de 8,6% par rapport à GPL. Toutefois, à partir du Tableau 5, l'approche avec la stratégie globale d'échantillonnage négative aléatoire échoue, tandis que l'approche proposée avec les deux autres stratégies d'échantillonnage négatives améliore respectivement le modèle de recherche dense D-BERT et GPL.

Ces approches montrent que l'approche d'annotation automatique proposée peut aider les modèles de recherche dense à généraliser vers de nouveaux domaines, et que le choix de la stratégie d'échantillonnage négative est important pour cela.

RQ2 Quel est l'impact des différentes stratégies d'échantillonnage négatives et laquelle est la meilleure ?

Dans les trois tableaux, nous observons une tendance globalement ascendante des trois différentes stratégies d'échantillonnage négatives. Bien que la méthode négative globale aléatoire améliore les modèles de recherche dense sur FiQA et Robust04, elle échoue sur le jeu de données BioASQ, pour lequel la distinction entre documents pertinents et non pertinents semble plus difficile.

En ce qui concerne les stratégies d'échantillonnage de négatifs BM25 et SimANS, elles montrent de meilleures performances que la stratégie de négatifs aléatoires globaux, et améliorent D-BERT et GPL sur les trois ensembles de données. Ces résultats montrent que l'échantillonnage de négatifs difficiles est important pour l'annotation automatique en RI.

L'approche proposée avec l'échantillonnage de négatifs difficiles SimANS donne systématiquement les meilleurs résultats sur tous les ensembles de données, meilleure que les stratégies de négatifs aléatoires globaux et de négatifs BM25, montrant qu'une meilleure stratégie d'échantillonnage de

7. https://huggingface.co/GPL

négatifs peut également améliorer davantage l'approche proposée.

RQ3 Quel est l'effet de l'approche d'annotation proposée avec l'échantillonnage de négatifs SimANS par rapport aux modèles de référence ?

Nous analysons maintenant l'approche proposée avec la stratégie d'échantillonnage de négatifs SimANS en la comparant aux modèles de référence.

BM25 est un algorithme de recherche standard considéré comme un modèle sans apprentissage. Bien qu'il soit extrêmement simple par rapport aux approches neurales de RI récentes, c'est une baseline solide et performante sur de nouveaux domaines. Surtout sur BioASQ, l'approche BM25 surpasse même l'approche de réordonnancement BM25 + CE sur un nouveau domaine. Notre approche proposée le surpasse sur FiQA et Robust04. En particulier, DoDress-T53B (GPL) avec SimANS est la seule approche qui surpasse BM25 sur Robust04, où GPL et TSDAE + GPL échouent. Sur BioASQ, tous les modèles denses sont inférieurs à BM25, tandis que DoDress-T53B (GPL) est le meilleur parmi eux.

Pour UDALM, MoDIR (ANCE) et les trois approches basées sur le pré-entraînement SimCSE, ICT et TSDAE, sur Robust04 nous n'avons pas les points de contrôle entraînés pour évaluer les 150 dernières requêtes de test, et nous montrons les résultats sur FiQA et BioASQ. DoDress-T53B (D-BERT) et DoDress-T53B (GPL) avec des négatifs obtenus par SimANS surclassent systématiquement tous les autres modèles. La meilleure baseline sur FiQA est MoDIR (ANCE) avec 29,6, tandis que nos DoDress-T53B (D-BERT) et DoDress-T53B (GPL) sont respectivement à 31,0 et 34,9. Sur BioASQ, la meilleure baseline parmi ces modèles est TSDAE avec 55,5, tandis que nos DoDress-T53B (D-BERT) et DoDress-T53B (GPL) sont à 60,6 et 65,3 respectivement, le dernier ayant une amélioration de 17,7%.

Les modèles de génération sont des modèles d'état de l'art précédents, principalement basés sur les pseudo-requêtes générées à partir de documents considérés comme pertinents. Dans notre approche, nous prenons également en compte les vraies requêtes. Dans le Tableau 3, nous voyons que TSDAE + GPL est le meilleur des modèles de base, avec un score de 34,4, mieux que les 32,8 de GPL. L'approche proposée, DoDress-T53B (GPL), obtient 34,9, ce qui est plus élevé que ces scores. Cependant, sur Robust04, dans le Tableau 4, TSDAE + GPL est moins bon que GPL : 40,7 et 41,9 respectivement. Notre approche proposée avec des négatifs difficiles échantillonnés avec SimANS, DoDress-T53B (D-BERT) et DoDress-T53B (GPL), les surpasse tous les deux, montrant ainsi l'efficacité de l'approche. Sur le jeu de données BioASQ, DoDress-T53B (GPL) obtient de meilleurs résultats que le meilleur modèle GPL dans les modèles de base.

RQ4 L'échantillonnage de documents non pertinents dans la liste fournie par un modèle dense fonctionne-t-il mieux et SimANS peut-il encore l'améliorer ?

Nous voulons voir si échantillonner les négatifs dans la liste du modèle de recherche dense conduit à de meilleurs résultats que l'échantillonnage dans la liste BM25. Nous menons donc une expérience supplémentaire en utilisant un échantillonnage négatif aléatoire à partir de la liste de recherche du modèle dense que nous cherchons à construire. Les résultats sont présentés dans le Tableau 2. Nous pouvons voir que l'échantillonnage des négatifs à partir de la liste fournie par le modèle D-BERT en cours de création est meilleur que celui à partir de la liste BM25. Cela peut s'expliquer par le fait que les négatifs provenant de la liste D-BERT sont plus ambigus pour ce modèle et donc plus informatifs pour son entraînement et l'adaptation à un nouveau domaine.

En outre, nous voulons voir si l'échantillonnage de SimANS peut encore améliorer les résultats de

TABLE 2 – Résultats de DoDress-BM25 (D-BERT) sur Robust04 avec différentes sources d'échantillonnage aléatoire de négatifs.

Échantillonnage aléatoire à partir de	nDCG@10 (%)
Liste supérieure BM25	41.6
Liste supérieure GPL	42.6

recherche. Dans le tableau 4, nous constatons que DoDress-BM25 (D-BERT) utilisant l'approche d'échantillonnage SimANS obtient 43,6, tandis que dans le tableau 2, il est de 42,6, ce qui montre que SimANS peut encore améliorer le résultat en échantillonnant des négatifs plus ambigus que l'échantillonnage aléatoire à partir de la liste de classement supérieure du modèle dense actuel.

En conclusion, les résultats ci-dessus démontrent l'efficacité de l'approche proposée combinant différents modèles. Ils confirment également l'importance du choix des documents non pertinents et le bon comportement de l'approche SimANS dans ce cadre.

TABLE 3 – Résultat d'adaptation de domaine de FiQA (en utilisant uniquement les requêtes d'entraînement).

modèle	nDCG@10 (%)
Modèles sans adaptation	
D-BERT	26.7
BM25 (Anserini)	23.6
Re-Ranking avec des Cross-Encoders (limite supérieure)	
BM25 + CE	33.1
BM25 + T53B	39.2
Méthodes précédentes d'adaptation de domaine	
UDALM	23.3
MoDIR (ANCE)	29.6
Pré-entraînement basé : Cible → D-BERT	
SimCSE	26.7
ICT	27.0
TSDAE	29.3
Basé sur la génération (SOTA précédent)	
QGen	28.7
GPL	32.8
TSDAE + GPL	34.4
Proposée : T53B, Négatifs Aléatoires Globaux	
DoDress-T53B (D-BERT)	27.3
DoDress-T53B (GPL)	33.0
Proposé : T53B, Négatifs durs BM25	
DoDress-BM25 (D-BERT)	30.4
DoDress-BM25 (GPL)	34.2
Proposé : T53B, Négatifs durs SimANS	
DoDress-T53B (D-BERT)	**31.0**
DoDress-T53B (GPL)	**34.9**

6 Conclusion

Nous avons étudié dans cet article s'il est possible d'annoter automatiquement des documents dans un domaine cible de façon à y déployer un modèle de RI dense. Notre étude révèle que cette approche fonctionne bien lorsque les annotations sont générées à l'aide d'un modèle T53B ré-ordonnant les documents obtenus par BM25, et qu'elle aide à améliorer les résultats de généralisation du modèle GPL qui utilise également des requêtes générées et des documents pertinents associés sur la collection

TABLE 4 – Résultats d'adaptation de domaine de Robust04 (l'ensemble d'entraînement et de développement utilise les 100 premières requêtes, l'ensemble de test est constitué des 150 dernières requêtes).

modèle	nDCG@10 (%)
Modèles sans adaptation	
D-BERT	39.1
BM25 (Anserini)	44.4
Re-Ranking avec des Cross-Encoders (limite supérieure)	
BM25 + CE	45.8
BM25 + T53B	51.8
Basé sur la génération (SOTA précédent)	
GPL	41.9
TSDAE + GPL	40.7
Proposée : T53B, Négatifs Aléatoires Globaux	
DoDress-T53B (D-BERT)	40.5
DoDress-T53B (GPL)	43.2
Proposé : T53B, Négatifs durs BM25	
DoDress-BM25 (D-BERT)	41.6
DoDress-BM25 (GPL)	43.3
Proposé : T53B, Négatifs durs SimANS	
DoDress-T53B (D-BERT)	**43.6**
DoDress-T53B (GPL)	**45.5**

TABLE 5 – Résultat d'adaptation de domaine de BioASQ.

modèle	nDCG@10 (%)
Modèles sans adaptation	
D-BERT	53.6
BM25 (Anserini)	73.0
Re-Ranking avec des Cross-Encoders (limite supérieure)	
BM25 + CE	72.8
BM25 + T53B	76.1
Méthodes précédentes d'adaptation de domaine	
UDALM	33.1
MoDIR (ANCE)	47.9
Pré-entraînement basé : Cible → D-BERT	
SimCSE	53.2
ICT	55.3
TSDAE	55.5
Basé sur la génération (SOTA précédent)	
QGen	56.5
GPL	62.8
TSDAE + GPL	61.6
Proposée : T53B, Négatifs Aléatoires Globaux	
DoDress-T53B (D-BERT)	52.9
DoDress-T53B (GPL)	62.0
Proposé : T53B, Négatifs durs BM25	
DoDress-BM25 (D-BERT)	58.6
DoDress-BM25 (GPL)	64.7
Proposé : T53B, Négatifs durs SimANS	
DoDress-T53B (D-BERT)	**60.6**
DoDress-T53B (GPL)	**65.3**

cible.

Nous avons également étudié l'importance du choix de la stratégie d'échantillonnage de documents non pertinents. Les meilleurs résultats sont obtenus en utilisant SimANS, une stratégie récente d'échantillonnage de documents non pertinents à partir des listes de documents obtenues par le modèle dense que l'on cherche à déployer dans le domaine cible.

Remerciements

Ce travail a été partiellement financé par MIAI@Grenoble Alpes (ANR-19-P3IA-0003) et la bourse du Chinese Scholarship Council (CSC) numéro 201906960018.

Références

[1] BLANCHARD G., DESHMUKH A. A., DOGAN Ü., LEE G. & SCOTT C. (2021). Domain generalization by marginal transfer learning. *The Journal of Machine Learning Research*, **22**(1), 46–100.

[2] BURGES C. J. (2010). From ranknet to lambdarank to lambdamart : An overview. *MSR-Tech Report*.

[3] CARLUCCI F. M., D'INNOCENTE A., BUCCI S., CAPUTO B. & TOMMASI T. (2019). Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, p. 2229–2238.

[4] DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). Bert : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186.

[5] DU Y., XU J., XIONG H., QIU Q., ZHEN X., SNOEK C. G. & SHAO L. (2020). Learning to learn with variational information bottleneck for domain generalization. In *European Conference on Computer Vision*, p. 200–216 : Springer.

[6] D'INNOCENTE A. & CAPUTO B. (2018). Domain generalization with domain-specific aggregation modules. In *German Conference on Pattern Recognition*, p. 187–198 : Springer.

[7] GANIN Y., USTINOVA E., AJAKAN H., GERMAIN P., LAROCHELLE H., LAVIOLETTE F., MARCHAND M. & LEMPITSKY V. (2016). Domain-adversarial training of neural networks. *The journal of machine learning research*, **17**(1), 2096–2030.

[8] GAO L. & CALLAN J. (2021). Condenser : a pre-training architecture for dense retrieval. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 981–993.

[9] GAO L. & CALLAN J. (2022). Unsupervised corpus aware language model pre-training for dense passage retrieval. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 2843–2853.

[10] GAO T., YAO X. & CHEN D. (2021). Simcse : Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 6894–6910.

[11] GUO J., FAN Y., PANG L., YANG L., AI Q., ZAMANI H., WU C., CROFT W. B. & CHENG X. (2020). A deep look into neural ranking models for information retrieval. *Information Processing & Management*, **57**(6), 102067.

[12] HOFSTÄTTER S., ALTHAMMER S., SCHRÖDER M., SERTKAN M. & HANBURY A. (2020). Improving efficient neural ranking models with cross-architecture knowledge distillation. *arXiv preprint arXiv :2010.02666*.

[13] JEON S., HONG K., LEE P., LEE J. & BYUN H. (2021). Feature stylization and domain-aware contrastive learning for domain generalization. In *Proceedings of the 29th ACM International Conference on Multimedia*, p. 22–31.

[14] KAROUZOS C., PARASKEVOPOULOS G. & POTAMIANOS A. (2021). Udalm : Unsupervised domain adaptation through language modeling. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 2579–2590.

[15] KARPUKHIN V., OGUZ B., MIN S., LEWIS P., WU L., EDUNOV S., CHEN D. & YIH W.-T. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 6769–6781.

[16] KUN ZHOU, YEYUN GONG X. L. W. X. Z. Y. S. A. D. J. L. R. M. J.-R. W. N. D. & CHEN W. (2022). Simans : Simple ambiguous negatives sampling for dense text retrieval.

[Laignelet & Rioult] LAIGNELET M. & RIOULT F. Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs.

[Langlais & Patry] LANGLAIS P. & PATRY A. Enrichissement d'un lexique bilingue par analogie. p. 101–110.

[19] LEE K., CHANG M.-W. & TOUTANOVA K. (2019). Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 6086–6096.

[20] LI D., YANG Y., SONG Y.-Z. & HOSPEDALES T. (2018). Learning to generalize : Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

[21] LI D., YANG Y., SONG Y.-Z. & HOSPEDALES T. M. (2017). Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, p. 5542–5550.

[22] LI M. & GAUSSIER E. (2022). Bert-based dense intra-ranking and contextualized late interaction via multi-task learning for long document retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 2347–2352.

[23] LIANG D., XU P., SHAKERI S., SANTOS C. N. D., NALLAPATI R., HUANG Z. & XIANG B. (2020). Embedding-based zero-shot retrieval through query generation. *arXiv preprint arXiv :2009.10270*.

[24] LIU C., SUN X., WANG J., TANG H., LI T., QIN T., CHEN W. & LIU T.-Y. (2021). Learning causal semantic representation for out-of-distribution prediction. *Advances in Neural Information Processing Systems*, **34**, 6155–6170.

[25] LOSHCHILOV I. & HUTTER F. (2017). SGDR : Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*.

[26] MA J., KOROTKOV I., YANG Y., HALL K. & MCDONALD R. (2021). Zero-shot neural passage retrieval via domain-targeted synthetic question generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : Main Volume*, p. 1075–1088.

[27] MAIA M., HANDSCHUH S., FREITAS A., DAVIS B., MCDERMOTT R., ZARROUK M. & BALAHUR A. (2018). Www'18 open challenge : Financial opinion mining and question answering. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, p. 1941–1942, Republic and Canton of Geneva, CHE : International World Wide Web Conferences Steering Committee. DOI : 10.1145/3184558.3192301.

[28] MANCINI M., BULO S. R., CAPUTO B. & RICCI E. (2018). Best sources forward : domain generalization through source-specific nets. In *2018 25th IEEE international conference on image processing (ICIP)*, p. 1353–1357 : IEEE.

[29] MICIKEVICIUS P., NARANG S., ALBEN J., DIAMOS G., ELSEN E., GARCIA D., GINSBURG B., HOUSTON M., KUCHAIEV O., VENKATESH G. *et al.* (2018). Mixed precision training. In *International Conference on Learning Representations*.

[30] MOKRII I., BOYTSOV L. & BRASLAVSKI P. (2021). A systematic evaluation of transfer learning and pseudo-labeling with bert-based ranking models. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 2081–2085.

[31] MOTIIAN S., PICCIRILLI M., ADJEROH D. A. & DORETTO G. (2017). Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE international conference on computer vision*, p. 5715–5725.

[32] NAM H., LEE H., PARK J., YOON W. & YOO D. (2021). Reducing domain gap by reducing style bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, p. 8690–8699.

[33] NGUYEN T., ROSENBERG M., SONG X., GAO J., TIWARY S., MAJUMDER R. & DENG L. (2016). MS MARCO : A human generated machine reading comprehension dataset. In T. R. BESOLD, A. BORDES, A. S. D'AVILA GARCEZ & G. WAYNE, Éds., *Proceedings of the Workshop on Cognitive Computation : Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*, volume 1773 de *CEUR Workshop Proceedings* : CEUR-WS.org.

[34] 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.

[35] PRAKASH A., BOOCHOON S., BROPHY M., ACUNA D., CAMERACCI E., STATE G., SHAPIRA O. & BIRCHFIELD S. (2019). Structured domain randomization : Bridging the reality gap by context-aware synthetic data. In *2019 International Conference on Robotics and Automation (ICRA)*, p. 7249–7255 : IEEE.

[36] QIAO F., ZHAO L. & PENG X. (2020). Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, p. 12556–12565.

[37] RAFFEL C., SHAZEER N., ROBERTS A., LEE K., NARANG S., MATENA M., ZHOU Y., LI W., LIU P. J. *et al.* (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, **21**(140), 1–67.

[38] RAHMAN M. M., FOOKES C., BAKTASHMOTLAGH M. & SRIDHARAN S. (2019). Multi-component image translation for deep domain generalization. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, p. 579–588 : IEEE.

[39] 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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 3982–3992.

[40] ROBERTSON S. E. & ZARAGOZA H. (2009). The probabilistic relevance framework : BM25 and beyond. *Found. Trends Inf. Retr.*, **3**(4), 333–389.

[41] SANH V., DEBUT L., CHAUMOND J. & WOLF T. (2019). Distilbert, a distilled version of BERT : smaller, faster, cheaper and lighter. *arXiv preprint arXiv :1910.01108*.

[Seretan & Wehrli] SERETAN V. & WEHRLI E. Collocation translation based on sentence alignment and parsing. p. 401–410.

[43] SHANKAR S., PIRATLA V., CHAKRABARTI S., CHAUDHURI S., JYOTHI P. & SARAWAGI S. (2018). Generalizing across domains via cross-gradient training. In *International Conference on Learning Representations*.

[44] 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)*.

[45] TOBIN J., FONG R., RAY A., SCHNEIDER J., ZAREMBA W. & ABBEEL P. (2017). Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, p. 23–30 : IEEE.

[46] TSATSARONIS G., BALIKAS G., MALAKASIOTIS P., PARTALAS I., ZSCHUNKE M., ALVERS M. R., WEISSENBORN D., KRITHARA A., PETRIDIS S., POLYCHRONOPOULOS D. *et al.* (2015). An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, **16**(1), 138.

[47] VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł. & POLOSUKHIN I. (2017). Attention is all you need. *Advances in neural information processing systems*, **30**.

[48] VOLPI R., NAMKOONG H., SENER O., DUCHI J. C., MURINO V. & SAVARESE S. (2018). Generalizing to unseen domains via adversarial data augmentation. *Advances in neural information processing systems*, **31**.

[49] VOORHEES E. (2005). : Special Publication (NIST SP), National Institute of Standards and Technology, Gaithersburg, MD.

[50] WANG J., LAN C., LIU C., OUYANG Y., QIN T., LU W., CHEN Y., ZENG W. & YU P. (2022a). Generalizing to unseen domains : A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*.

[51] WANG K., REIMERS N. & GUREVYCH I. (2021). Tsdae : Using transformer-based sequential denoising auto-encoderfor unsupervised sentence embedding learning. In *Findings of the Association for Computational Linguistics : EMNLP 2021*, p. 671–688.

[52] WANG K., THAKUR N., REIMERS N. & GUREVYCH I. (2022b). GPL : Generative pseudo labeling for unsupervised domain adaptation of dense retrieval. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics : Human*

Language Technologies, p. 2345–2360, Seattle, United States : Association for Computational Linguistics. DOI : 10.18653/v1/2022.naacl-main.168.

[53] WANG M. & DENG W. (2018). Deep visual domain adaptation : a survey. *Neurocomputing*.

[54] XIN J., XIONG C., SRINIVASAN A., SHARMA A., JOSE D. & BENNETT P. (2022). Zero-shot dense retrieval with momentum adversarial domain invariant representations. In *Findings of the Association for Computational Linguistics : ACL 2022*, p. 4008–4020.

[55] XIONG L., XIONG C., LI Y., TANG K.-F., LIU J., BENNETT P. N., AHMED J. & OVERWIJK A. (2020). Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*.

[56] YANG P., FANG H. & LIN J. (2018). Anserini : Reproducible ranking baselines using lucene. *Journal of Data and Information Quality (JDIQ)*, **10**(4), 1–20.

[57] ZHANG H., CISSE M., DAUPHIN Y. N. & LOPEZ-PAZ D. (2018). mixup : Beyond empirical risk minimization. In *International Conference on Learning Representations*.