Actes de CORIA-TALN-RJCRI-RECITAL 2025
Marseille, France, 30 juin au 4 juillet 2025

Comprendre la Nature des Signaux de Correspondance dans les Modèles Neuronaux pour la RI

Mathias Vast[1, 2] Basile Van Cooten[1] Laure Soulier[2] Benjamin Piwowarski[2]
(1) Sinequa by ChapsVision, Paris, France
(2) Sorbonne Université, CNRS, ISIR, F-75005 Paris, France
{prenom.nom}@isir.upmc.fr, bvancooten@chapsvision.com,

RÉSUMÉ

Les architectures de recherche d'information (RI) neuronale, en particulier celles basées sur l'interaction, sont des modèles très performants dont les mécanismes restent largement méconnus. La plupart des travaux visant à expliquer leur comportement se sont attachés à décrire des processus en surface (par exemple, quels éléments de l'entrée influencent la prédiction ? le modèle respecte t-il les axiomes connus de la RI ?) mais ne décrivent pas précisément le processus d'appariement. Dans cet article, nous apportons un nouvel éclairage sur le mécanisme de correspondance en analysant le processus d'attention, et en mettant en évidence le rôle crucial de certaines têtes d'attention ainsi que la nature des signaux qui sont manipulés. [1]

ABSTRACT

On the Nature of Matching Signals in Neural IR Models

Neural Information Retrieval architectures, particularly cross-encoders, are highly effective models whose internal mechanisms are mostly unknown. Most works trying to explain their behavior focused on high-level processes (e.g., what in the input influences the prediction, does the model adhere to known IR axioms) but fall short of describing the matching process. In this paper, we focus on the attention process and extract causal insights highlighting the crucial roles of some attention heads in this process as well as the nature of these matching signals.

MOTS-CLÉS : Interprétabilité, Approches basées sur l'interaction, Attention, Pertinence.

KEYWORDS: Interpretability, Cross-Encoders, Attention, Relevance.

1 Introduction

Avec l'essor des modèles basés sur les Transformeurs (31), culminant actuellement avec la démocratisation des grands modèles de langage (LLM), la recherche d'information (RI) a radicalement changé depuis l'ère BM25 (1). Encodant les concepts et la sémantique en vecteurs, les Transformeurs ont permis de dépasser les approches par mots-clefs. Trois grands types de modèles se distinguent actuellement, avec différents compromis efficacité-efficience : approches neuronales basées sur l'interaction forte (25; 26), sur l'interaction faible (20; 15) ou sur la représentation (19) selon la façon dont sont encodées les requêtes et les documents. Malgré leur succès, ces modèles sont considérés comme des boîtes noires et de nombreuses zones d'ombre entourent leur fonctionnement interne. Cette lacune limite les améliorations telles que l'adaptation à de nouvelles données ou

1. Le code est disponible ici : https://git.isir.upmc.fr/mat_vast/sigir25-matching-signals

la simplification de couches de complexité superflues. La plupart des travaux d'explicabilité en RI visent à identifier les éléments d'entrée ou les composants du modèle qui contribuent le plus aux prédictions (2), renforçant la transparence et la confiance dans le modèle et laissant entrevoir ses limites. Par exemple, (30) révèle l'existence de deux ensembles de neurones dans MonoBERT (25), dédiés respectivement à la prédiction de la pertinence et des situations hors domaine, tandis que (24) met en évidence l'importance des noms et des verbes dans la prédiction d'encodeurs siamois. Une autre direction consiste à examiner si les modèles basés sur les Transformeurs adhèrent aux axiomes traditionnels de la RI (14). Dans le cas des approches basées sur l'interaction telles que MonoBERT, cela a permis de montrer leurs capacités d'appariement syntaxique mais aussi de mieux mesurer leurs capacités sémantiques (4). Parallèlement, les études de « sondage »(3) mettent l'accent sur la capacité des modèles basés sur BERT (13) à retrouver des mots-clés mais également sur la capacité des premières couches à traiter et capturer la similarité sémantique d'un document avec une requête (33). Même si les résultats des études de « probing »restent des corrélations et ne permettent pas à eux seuls d'expliquer le fonctionnement des modèles, ils fournissent des indications précieuses sur les propriétés de base utilisées en interne par les modèles. L'analyse des couches intermédiaires révèle que les modèles de RI apprennent implicitement à effectuer des tâches étroitement liées à la RI, telles que la reconnaissance d'entités nommées et la résolution de coréférences (29; 33). Ces modèles encodent aussi des caractéristiques sur les statistiques liées à la requête et au document, telles que les scores tf-idf ou le nombre de termes de la requête couverts, connus pour être importants pour les modèles d'ordonnancement (Choi *et al.*; 7). Si ces découvertes concernent principalement les modèles d'interaction forte, des études parallèles (16; 21) ont également fait des découvertes similaires sur des modèles d'interaction tardive tels que ColBERT (20).

Malgré l'importance de leurs apports à la communauté, ces travaux ne clarifient pas entièrement les rôles des composants du modèle ou les processus sous-jacents à la prédiction de la pertinence. Des études antérieures ont permis de réaliser des percées dans cette direction dans le cadre de la RI. Zhan *et al.* (34) étudient le processus d'attention (31) de MonoBERT. Ils montrent que bien que les tokens `CLS` et `SEP` et la ponctuation attirent une grande partie de l'attention, ils ne contiennent que peu d'informations pertinentes, ce qui fait écho au comportement de « no-op(eration) » (8). Ils soulignent de plus l'importance des transferts d'information de la requête vers le document tout en remettant en question le rôle des transferts en sens inverse. Leurs conclusions suggèrent aussi que MonoBERT prédit la pertinence en contextualisant d'abord les requêtes et les documents, puis en capturant les signaux d'interaction et enfin en les intégrant à la prédiction. Dans l'ensemble, leur étude fournit diverses informations sur les mécanismes internes de MonoBERT et ouvre la voie à des améliorations potentielles. Cependant, ils ne précisent pas les types d'interactions détectés ni la manière dont elles contribuent au signal de pertinence, et leurs conclusions sur la direction de l'attention manquent de preuves empiriques plus solides. Il est intéressant de noter que Chen *et al.* (5) a découvert les rôles clés de certaines têtes d'attention dans la détection des copies des mêmes tokens et dans la composition du signal de pertinence en étudiant TAS-B (17), un bi-encodeur, en utilisant une méthode d'intervention causale (23; 32). Ces résultats font écho au processus décrit dans (34).

Dans cet article, nous étendons (34) et approfondissons la façon dont les modèles neuronaux de RI construisent le signal de pertinence. Plus particulièrement, nous nous concentrons sur les interactions entre les requêtes et les documents en analysant le flux d'information dans le modèle. Pour ce faire, nous étudions les architectures basées sur l'interaction car cette dernière modélise directement les interactions entre les requêtes et les documents, avec le même niveau de granularité que dans

Chen *et al.* (5). Guidés par les conclusions similaires obtenues dans (5; 34) suggérant que certains mécanismes dans les modèles neuronaux de RI pourraient être agnostiques en termes d'architecture, nous pensons que notre travail peut révéler des mécanismes généralisables à l'ensemble des modèles neuronaux de RI. Nos contributions sont les suivantes : **(C1)** Nous étendons l'étude d'ablation de (34) et validons l'importance de la contextualisation du document et des transferts d'informations de la requête vers le document. Toutefois, nous remettons en question leur conclusion sur l'importance des transferts du document vers la requête. Nous confirmons le rôle du `SEP` dans l'opération « no-op » et montrons que le `CLS` pourrait jouer un rôle similaire, des premières couches jusqu'aux couches intermédiaires. **(C2)** En contrastant les motifs d'attention entre la requête et les passages pertinents versus des passages non pertinents, nous découvrons des têtes d'attention spécialisées dans la détection des interactions liées à la pertinence. En outre, nous montrons que, des premières couches aux couches intermédiaires, ces interactions sont des correspondances entre les termes de la requête et du document (*correspondance syntaxique*), et des correspondances entre des tokens sémantiquement contextualisés (*correspondance sémantique*) pour les dernières couches.

2 Protocole Expérimental

Nous présentons ici les ensembles de données utilisés ainsi que les notations suivies pour désigner les différentes parties des matrices d'attention. Pour ce travail, nous utilisons une version de MonoBERT-large publique et déjà entraînée sur la tâche de RI [2]. L'indexation des couches démarre à 0.

2.1 Jeux de données

Dans nos expériences, nous nous appuyons sur les jeux de données des tracks « TREC Deep-Learning » de 2019 à 2022 (9; 10; 11; 12), accessibles via la librairie *ir-datasets* (22), qui contiennent plusieurs passages annotés par requête [3]. Nous nous limitons aux requêtes qui ont été annotées manuellement par les évaluateurs du NIST (version *judged* des données sur *ir-datasets*) pour un total de 226 requêtes et plus de 400k jugements de pertinence, décliné selon 4 niveaux $r \in \{0..3\}$, à travers les 4 ensembles de données.

Comme les documents étiquetés comme « non pertinents » présentent quand même une proximité syntaxique avec la requête (un moteur de recherche de type BM25 les a identifiés), nous étudions également le comportement de MonoBERT dans le cas où le document est non pertinent *et* non syntaxiquement lié, en échantillonnant aléatoirement des documents dans l'ensemble de données. Conformément à la terminologie utilisée dans la littérature de la RI neuronale, nous appelons ces documents *négatifs faciles* (où $r = -1$) par opposition aux *négatifs difficiles* (28) (où $r = 0$).

2.2 Analyse des matrices d'Attention

Comme cette étude se concentre sur les mécanismes en jeu au niveau de l'Attention (31), nous étendons Zhan *et al.* (34) et étudions les *directions* suivies par l'information à travers les parties de l'entrée. Pour chaque paire requête-document, nous considérons 5 parties d'entrée distinctes : le token CLS, la requête Q, le premier token $SEP1$, le document D et le dernier token $SEP2$. Dans le processus d'attention, chacune de ces parties de l'entrée peut porter attention à elle-même

2. Le modèle est disponible sur le hub de HuggingFace : castorini/monobert-large-msmarco
3. Notez que dans *ir-datasets*, l'ensemble de documents TREC-DL 2022 a été dédupliqué.

ou à n'importe laquelle des autres, ce qui donne 25 *directions* possibles. Nous appelons *direction*, l'attention de la partie d'entrée X à la partie d'entrée Y dans la matrice de poids de l'attention ou, de manière équivalente, l'information transmise par la partie d'entrée Y à la partie d'entrée X, dénotée $X \leftarrow Y$ avec X, Y dans $\{CLS, Q, SEP1, D, SEP2\}$. Pour la $h$ème tête d'attention de la couche ℓ, nous notons $p_{h,\ell}(X_i \leftarrow Y_j)$ la probabilité que X_i, le i^{me} token de X, *fasse attention à* Y_j, le j^{me} token de Y.

3 Expérimentations et Résultats

Nous présentons 2 études menées dans ce contexte. Dans la section 3.1, nous détaillons l'étude d'ablation, puis décrivons dans la section 3.2 l'approche contrastive utilisée afin d'en apprendre plus sur la nature des signaux d'appariement que le modèle capture.

3.1 Etude d'ablations

Pour identifier les types d'attention les plus importants, nous forçons le modèle à les ignorer et quantifions les perturbations que cela introduit sur la tâche de RI. En pratique, nous perturbons l'inférence en masquant des *directions* dans la matrice d'attention (voir la section 2.2). Par exemple, $Q \nleftarrow D$ signifie que l'entrée correspondant à l'attention $p_{h,\ell}(Q_i \leftarrow D_j)$ est forcée à 0, quelle que soit sa valeur réelle. Zhan *et al.* (34) appellent cette stratégie de masquage « *attention mask* ». Contrairement à eux, notre étude d'ablation inclut l'ensemble des directions décrites dans la section 2.2. Pour des raisons de temps de calcul, nous limitons notre configuration de RI à l'ensemble de données de la tâche TREC-DL 19 (12) et échantillonnons jusqu'à 20 passages par niveau de pertinence et par requête. Sur cette configuration, nous rapportons un score de 0.81 en nDCG@10 pour le modèle de base et de 0.48 en moyenne sur 1000 classements générés aléatoirement.

La manière la plus directe de quantifier l'importance d'une direction $X \leftarrow Y$, $X, Y \in \{CLS, Q, SEP1, D, SEP2\}$ pour la prédiction de la pertinence est de la masquer pour chaque tête d'attention h et chaque couche ℓ de MonoBERT. Les résultats de cette expérience sont présentés dans le tableau 1. Comme supposé dans (34), l'impact le plus significatif correspond à l'ablation de la direction $D \leftarrow D$ (0.67 vs 0.81 pour la performance originale), confirmant l'importance de la contextualisation du document. Il est frappant de constater que $CLS \not\leftarrow Q$ ou de $CLS \not\leftarrow D$ (pris séparemment) ne nuisent pas de manière significative aux performances (resp. 0,79 et 0,81 vs 0,81 pour la perf. originale ; avec une valeur p > 0,05). Notre résultat contredit donc (34), puisque les représentations des tokens du document stockent aussi un signal de pertinence fort - et pas seulement celles des tokens de la requête. Cela suggère en outre que les transferts d'information entre la requête et le document ne sont pas unidirectionnels : ils se produisent dans les deux directions $Q \leftarrow D$ et $D \leftarrow Q$. Néanmoins, nous notons $CLS \not\leftarrow Q$ provoque plus de perturbations que $CLS \not\leftarrow D$ (0.79 vs 0.81), ce qui pourrait indiquer que les représentations des tokens de la requête stockent des signaux de pertinence plus forts.

Pour mieux comprendre quelle partie de l'entrée est la principale cible des transferts, nous indiquons dans la colonne « Tous » du tableau 1 les résultats du masquage simultané de toutes les directions pointant vers la même partie de l'entrée, par exemple en masquant $Q \leftarrow \{CLS, Q, SEP1, D, SEP2\}$. Si l'ablation de toutes les directions conduisant au `CLS` rend logiquement le modèle aléatoire, les autres ablations soulignent également l'importance de la requête et du document dans le processus. En outre, elles soutiennent l'hypothèse selon laquelle le modèle utilise principalement les tokens `SEP`

TABLE 1 – Résultats de l'ablation en termes de nDCG@10. Les valeurs en **gras** indique une baisse par rapport à la performance original du modèle et * précise si la baisse est statistiquement significative ($p \leq 0.05$) selon le t-test de Student.

$\nleftarrow$	CLS	Query	SEP1	Doc.	SEP2	Tous
CLS	0.82	**0.79**	0.81	0.81	0.81	**0.48***
Query	0.82	**0.78**	0.81	**0.8**	0.81	**0.66***
SEP1	0.81	0.81	0.81	0.81	0.81	0.81
Doc.	0.81	**0.79**	0.81	**0.67***	0.82	**0.62***
SEP2	0.81	0.81	0.81	0.81	0.81	0.81

Perf. originale = 0.81 / Classement aléatoire = 0.48 (nDCG@10).

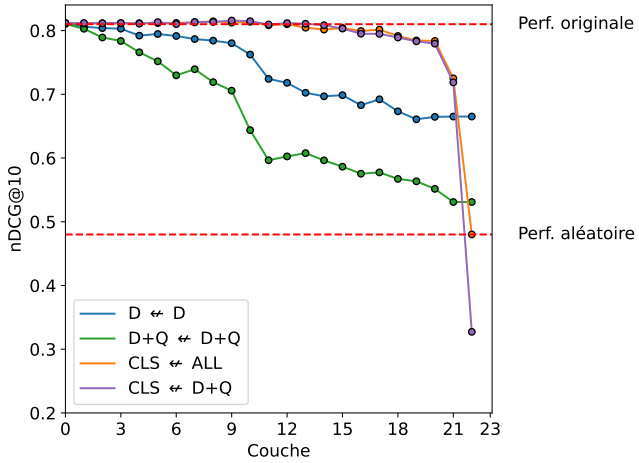


FIGURE 1 – Evolution du nDCG@10 suite à l'ablation de direction jusqu'à chaque couche.

comme une opération « no-op » (5; 34) puisque le fait de les masquer ne produit pas de perturbation.

Comme le souligne (34), la prédiction de la pertinence est un processus en plusieurs étapes, et le fait de masquer l'attention sur l'ensemble du modèle empêche de comprendre plus finement ce processus. Pour y remédier, nous masquons les directions $X \leftarrow Y$ jusqu'à chaque couche de MonoBERT, avec $X, Y \in \{CLS, Q, D\}$, car nous avons identifié ces parties comme étant les plus importantes. Dans la figure 1, nous observons que les ablations $CLS \nleftarrow \{Q, D\}$ et $CLS \nleftarrow \{CLS, Q, SEP1, D, SEP2\}$ ne nuisent pas à la performance avant les 2 dernières couches. Cela suggère que MonoBERT intègre le signal de pertinence dans la représentation des tokens CLS à la toute fin du processus. Avant cela, la manière dont le token CLS est utilisé n'est pas claire, mais les niveaux de perturbation suggèrent une fonction proche de celle des tokens SEP. De même, nous observons que le fait d'empêcher la contextualisation du document ne nuit pas aux performances avant les couches 11/12. Couplé à la présence de « têtes d'appariement » avant ces couches (voir la section 3.2), ces deux faits contredisent l'hypothèse selon laquelle le modèle contextualiserait d'abord le document avant de capturer les signaux d'interaction avec la requête (34). Au contraire, la contextualisation semble être précédée d'une autre étape, qui ne nécessite pas de contextualisation sémantique. Bien que nous ne puissions pas affirmer sa fonction, des études antérieures (4; 18; 33) suggèrent que c'est à ce niveau (couches précoces à intermédiaires) que (Mono)BERT capture des informations syntaxiques (*mots-clés*). À partir des couches où la contextualisation commence à se produire (couches intermédiaires à finales), le modèle capturerait donc également des informations sémantiques (*correspondance entre des tokens sémantiquement contextualisés*).

3.2 Analyse des motifs d'Attention

Pour approfondir la manière dont MonoBERT capture les signaux d'appariement lorsque les documents sont pertinents par rapport à la requête, nous utilisons une approche contrastive afin de faire ressortir d'éventuelles différences dans les motifs d'attention des têtes d'attention. En pratique, pour chaque tête d'attention h à la couche ℓ, pour un niveau de pertinence r et un token de la requête Q_i, nous calculons la probabilité d'attention maximale $\max_j p_{h,\ell}(Q_i \leftarrow D_j)$ sur un token du document D_j. Nous calculons ensuite la moyenne pour une requête donnée et désignons cette valeur par $M^{h,\ell}_{Q,D/r}$. Nous définissons de même $M^{h,\ell}_{D,Q/r}$ pour l'attention des tokens du document sur les tokens de la requête. Pour obtenir le plus de contraste possible, nous ne considérons que 3 niveaux de pertinence sur les 5 mentionnés précédemment : « pertinent » ($r = 3$), "négatif difficile" ($r = 0$), et "négatif facile" ($r = -1$), et comparons les différences, pour chaque tête d'attention, entre $r = 3$ et $r = -1$ ainsi qu'entre $r = 0$ et $r = -1$, à savoir : $M^{h,\ell}_{Q,D/3} - M^{h,\ell}_{Q,D/-1}$ et $M^{h,\ell}_{Q,D/0} - M^{h,\ell}_{Q,D/-1}$ resp.

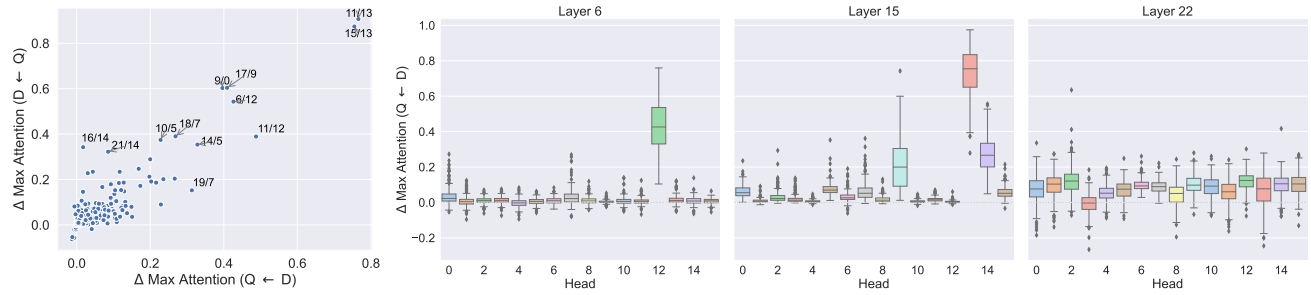
La combinaison des résultats de ces deux comparaisons permet d'identifier des comportements correspondant à la détection de signaux de correspondance et la nature de ces signaux. En effet, étant donné l'origine des négatifs difficiles, nous nous attendons à trouver des motifs similaires en comparant $r = 3$ à $r = -1$ et $r = 0$ à $r = -1$ lorsque les signaux de correspondance capturés sont syntaxiques, mais des motifs différents lorsqu'ils sont sémantiques. Les Figures 2 et 3 résument ces différences. Dedans, chaque tête du modèle est désignée par « index de la couche/index de la tête ».

Dans la figure 2a, on peut facilement repérer les têtes d'attention de MonoBERT spécialisées dans la capture des signaux de correspondance aux vues des fortes différences qu'elles présentent dans leurs motifs d'attention lorsque le document est pertinent ou non (par exemple, couche/tête 11/13). Nous observons également que ces têtes d'appariement sont *bi-directionnelles*, c'est-à-dire que les tokens de la requête tentent de matcher avec un token du document, et vice-versa, au niveau d'une *même tête d'attention*. Ceci confirme le fait que les ablations $D \leftarrow Q$ ou $Q \leftarrow D$ ne détruisent pas à elles seules le signal de pertinence (Section 3.1).

En examinant la distribution de ces têtes à travers les couches du modèle dans la Figure 2b, nous observons que des premières couches aux couches intermédiaires, seules quelques têtes par couche présentent une différence, contrairement aux dernières couches où chaque tête présente une différence (bien que plus petite en magnitude). Chen *et al.* (5) rapporte des observations similaires sur les têtes d'attention des couches intermédiaires de TAS-B (17).

La comparaison des passages pertinents aux négatifs difficiles ($r = 0$) confirme ces observations, même si les différences sont plus faibles en raison de la similarité syntaxique des négatifs difficiles avec les requêtes par rapport aux passages échantillonnés de manière aléatoire (non rapportés ici).

Comme nous ne pouvons pas encore différencier la nature des interactions capturées par chaque tête, nous examinons les différences entre $r = 0$ et $r = -1$ dans la Figure 3. Comme mentionné, les négatifs difficiles présentent une plus grande similarité syntaxique avec les requêtes que les négatifs faciles, tout en n'étant pas pertinents. Pour cette raison, nous nous attendons à répéter les mêmes observations qu'entre $r = 3$ et $r = -1$ pour les têtes capturant les signaux de *correspondance syntaxique* (fortes différences), tandis que les têtes chargées de capturer les signaux de *correspondance sémantique* devraient se comporter de la même manière. En effet, la Figure 3 confirme que pour $r = 0$ et $r = -1$, des têtes présentent une différence dans les couches initiales et intermédiaires (couches 6 et 15) alors que les dernières couches (couches 20 à 23) se comportent de manière similaire (médiane



(a) Moyenne de $M^{h,\ell}_{Q,D/3} - M^{h,\ell}_{Q,D/-1}$ vs $M^{h,\ell}_{D,Q/3} - M^{h,\ell}_{D,Q/-1}$ pour chaque tête d'attention (couche ℓ/tête h).

(b) Distribution des différences $M^{h,\ell}_{Q,D/3} - M^{h,\ell}_{Q,D/-1}$ (même requête).

FIGURE 2 – Différences dans les motifs d'attention entre pertinents et négatifs faciles ($r = 3$ et $r = -1$).

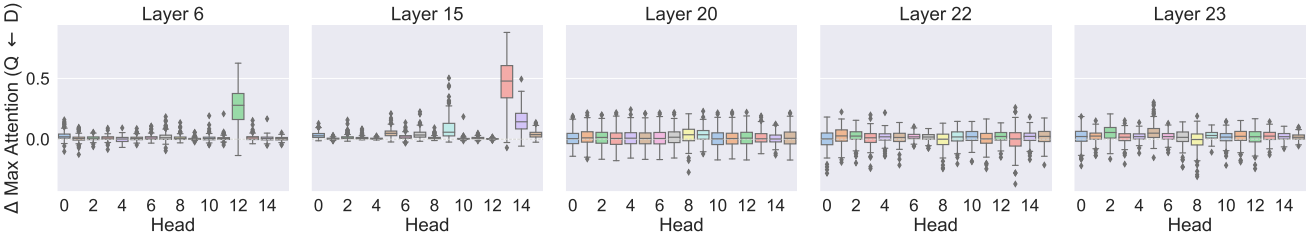


FIGURE 3 – Distribution des différences entre négatifs difficiles et faciles ($r = 0$ et $r = -1$) $M^{h,\ell}_{Q,D/0} - M^{h,\ell}_{Q,D/-1}$ (même requête).

proche de 0).

Pour renforcer cette conclusion, nous examinons les schémas d'attention lorsque la requête regarde le document (et vice-versa). Plus précisément, la Figure 4 montre les différences entre (1) l'entropie [4] de la distribution de l'attention $p_{h,\ell}(Q_i \leftarrow D_j | Q_i \leftarrow D)$ sur les tokens du document D_j, et (2) la probabilité $p_{h,\ell}(\bigvee_j Q_i \leftarrow D_j \wedge \hat{Q}_i = \hat{D}_j | Q_i \leftarrow D)$ portée par Q_i aux tokens qui lui sont identiques dans le document, avec $\hat{Q}_i$ et $\hat{D}_j$ resp. le i^{eme} token de la requête et le j^{eme} token du document. Dans les deux cas, nous calculons une moyenne de ces valeurs sur tous les termes de la requête, pondérée par la probabilité que le token de la requête soit associé au document $p_{h,\ell}(Q_i \leftarrow D)$. Nous observons que les têtes d'attention des dernières couches présentent une entropie plus élevée lorsque les passages sont pertinents (en raison de la contextualisation sémantique), contrairement aux têtes d'attention des premières couches et des couches intermédiaires qui ont tendance à s'intéresser davantage aux tokens identiques lorsque les passages sont pertinents, comme attendus.

Alors que les comportements observés dans les Figures 2 et 3 sont similaires pour les deux directions $D \leftarrow Q$ et $Q \leftarrow D$, nous notons que cela n'est pas vrai pour l'entropie des têtes dans les dernières couches dans la direction $D \leftarrow Q$. Contrairement à la direction $Q \leftarrow D$, les distributions pour les passages pertinents et non pertinents sont davantage similaires pour cette direction, ce qui signifie que l'appariement sémantique est unidirectionnel et va du document à la requête, contrairement

4. Entropie de Shannon (27) plus précisément, via l'implémentation Scipy.

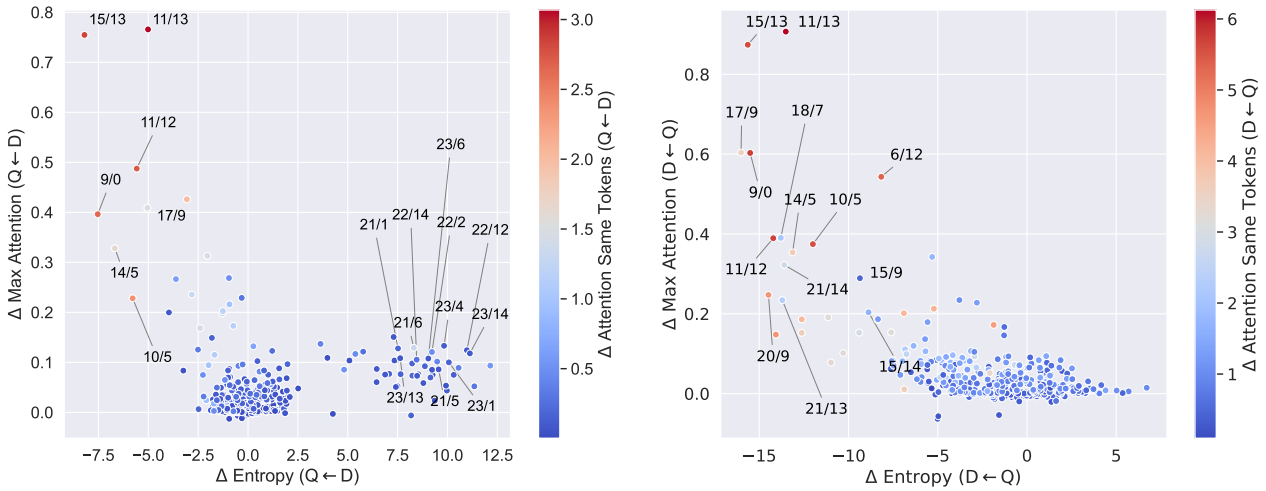


FIGURE 4 – Comparaison des distributions de l'attention pour $Q \leftarrow D$ et $D \leftarrow Q$ entre pertinents et négatifs faciles (couche ℓ/tête h).

à l'appariement syntaxique. Cela explique pourquoi nous avons observé plus de perturbations en masquant $CLS \leftarrow Q$ que $CLS \leftarrow D$ (voir Tableau 1), car le signal de pertinence stocké dans les représentations des tokens de la requête est plus fort.

4 Conclusion et Travaux Futurs

Grâce à ces expériences, nous avons montré que des approches simples, basées sur l'attention, peuvent être utilisées pour étudier les interactions requête-document dans les modèles neuronaux de RI tels que MonoBERT. Dans notre étude d'ablation, nous avons reproduit et étendu les résultats de Zhan *et al.* (34) confirmant l'importance de la contextualisation des documents, des interactions requête-document et de la fonction « no-op » des tokens `SEP`. En ce qui concerne le processus en plusieurs étapes menant à la prédiction de la pertinence, nous avons montré que l'étape de contextualisation du document et l'étape de capture des interactions entre la requête et le document sont imbriquées l'une dans l'autre et non pas consécutives. Nos observations suggèrent en outre que la composition du signal de pertinence a lieu dans les toutes dernières couches de MonoBERT. En opposant les motifs d'attention, nous avons démontré que MonoBERT s'appuie sur un nombre limité de têtes pour capturer les interactions requête-document. Dans les premières couches et celles intermédiaires, ces têtes capturent les signaux de correspondance syntaxique entre la requête et le document, tandis que dans les dernières couches, d'autres capturent les signaux de correspondance sémantique. Enfin, nous tenons à souligner que des travaux supplémentaires sont nécessaires pour comprendre pleinement le fonctionnement de ces modèles. Même si nous avons décrit le type de signaux d'appariement capturés par MonoBERT, le mécanisme en jeu demeure inconnu. Nous pensons que des approches inspirées de l'interprétabilité mécanistique (4) pourraient aider à résoudre ce problème.

Remerciements

Ce travail est soutenu par les projets ANR ANR-23-IAS1-0003 et ANR-23-IACL-0007.

Références

[1] AMATI G. & VAN RIJSBERGEN C. J. (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.*, **20**(4), 357–389. DOI : 10.1145/582415.582416.

[2] ANAND A., LYU L., IDAHL M., WANG Y., WALLAT J. & ZHANG Z. (2022). Explainable information retrieval : A survey. *arXiv preprint arXiv :2211.02405*.

[3] BELINKOV Y. (2022). Probing Classifiers : Promises, Shortcomings, and Advances. *Computational Linguistics*, **48**(1), 207–219. DOI : 10.1162/coli_a_00422.

[4] CÂMARA A. & HAUFF C. (2020). Diagnosing bert with retrieval heuristics. In J. M. JOSE, E. YILMAZ, J. MAGALHÃES, P. CASTELLS, N. FERRO, M. J. SILVA & F. MARTINS, Éds., *Advances in Information Retrieval*, p. 605–618, Cham : Springer International Publishing.

[5] CHEN C., MERULLO J. & EICKHOFF C. (2024). Axiomatic causal interventions for reverse engineering relevance computation in neural retrieval models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 1401–1410.

[Choi *et al.*] CHOI J., JUNG E., LIM S. & RHEE W. Finding Inverse Document Frequency Information in BERT. DOI : 10.48550/arXiv.2202.12191.

[7] CHOWDHURY T. & ALLAN J. (2024). Probing ranking llms : Mechanistic interpretability in information retrieval.

[8] CLARK K., KHANDELWAL U., LEVY O. & MANNING C. D. (2019). What does BERT look at ? an analysis of BERT's attention. In T. LINZEN, G. CHRUPAŁA, Y. BELINKOV & D. HUPKES, Éds., *Proceedings of the 2019 ACL Workshop BlackboxNLP : Analyzing and Interpreting Neural Networks for NLP*, p. 276–286, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/W19-4828.

[9] CRASWELL N., MITRA B., YILMAZ E. & CAMPOS D. (2021). Overview of the trec 2020 deep learning track. In *Text REtrieval Conference (TREC)* : TREC.

[10] CRASWELL N., MITRA B., YILMAZ E., CAMPOS D. & LIN J. (2022). Overview of the trec 2021 deep learning track. In *Text REtrieval Conference (TREC)* : NIST TREC.

[11] CRASWELL N., MITRA B., YILMAZ E., CAMPOS D., LIN J., VOORHEES E. M. & SOBOROFF I. (2023). Overview of the trec 2022 deep learning track. In *Text REtrieval Conference (TREC)* : NIST TREC.

[12] CRASWELL N., MITRA B., YILMAZ E., CAMPOS D. & VOORHEES E. M. (2020). Overview of the trec 2019 deep learning track. In *Text REtrieval Conference (TREC)* : TREC.

[13] 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* : Association for Computational Linguistics. DOI : 10.18653/v1/n19-1423.

[14] FANG H., TAO T. & ZHAI C. (2004). A formal study of information retrieval heuristics. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, p. 49–56, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/1008992.1009004.

[15] 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 : ACM. DOI : 10.1145/3404835.3463098.

[16] FORMAL T., PIWOWARSKI B. & CLINCHANT S. (2022). Match your words ! a study of lexical matching in neural information retrieval. In M. HAGEN, S. VERBERNE, C. MACDONALD, C. SEIFERT, K. BALOG, K. NØRVÅG & V. SETTY, Éds., *Advances in Information Retrieval*, p. 120–127, Cham : Springer International Publishing.

[17] HOFSTÄTTER S., LIN S.-C., YANG J.-H., LIN J. & HANBURY A. (2021). Efficiently teaching an effective dense retriever with balanced topic aware sampling. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 113–122.

[18] JAWAHAR G., SAGOT B. & SEDDAH D. (2019). What Does BERT Learn about the Structure of Language ? In A. KORHONEN, D. TRAUM & L. MÀRQUEZ, Éds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 3651–3657, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1356.

[19] KARPUKHIN V., OGUZ B., MIN S., LEWIS P., WU L., EDUNOV S., CHEN D. & YIH W.-T. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 6769–6781, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.550.

[20] KHATTAB O. & ZAHARIA M. (2020). Colbert : Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, p. 39–48, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3397271.3401075.

[21] MACAVANEY S., FELDMAN S., GOHARIAN N., DOWNEY D. & COHAN A. (2022). ABNIRML : Analyzing the behavior of neural IR models. *Transactions of the Association for Computational Linguistics*, **10**, 224–239. DOI : 10.1162/tacl_a_00457.

[22] MACAVANEY S., YATES A., FELDMAN S., DOWNEY D., COHAN A. & GOHARIAN N. (2021). Simplified data wrangling with ir$_datasets. InSIGIR$.

23] MENG K., BAU D., ANDONIAN A. & BELINKOV Y. (2022). Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, **35**, 17359–17372.

24] MOELLER L., NIKOLAEV D. & PADÓ S. (2023). An attribution method for Siamese encoders. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 15818–15827, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.emnlp-main.980.

25] NOGUEIRA R. & CHO K. (2019). Passage re-ranking with bert. *arXiv preprint arXiv :1901.04085*.

26] 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* : Association for Computational Linguistics. DOI : 10.18653/v1/2020.findings-emnlp.63.

27] SHANNON C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, **27**(3), 379–423. DOI : 10.1002/j.1538-7305.1948.tb01338.x.

28] TONELLOTTO N. (2022). Lecture Notes on Neural Information Retrieval.

29] VAN AKEN B., WINTER B., LÖSER A. & GERS F. A. (2019). How does bert answer questions ? : A layer-wise analysis of transformer representations. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19 : ACM. DOI : 10.1145/3357384.3358028.

30] VAST M., VAN COOTEN B., SOULIER L. & PIWOWARSKI B. (2024). Which neurons matter in ir ? applying integrated gradients-based methods to understand cross-encoders. In *Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '24, p. 133–143 : ACM. DOI : 10.1145/3664190.3672528.

31] VASWANI A., SHAZEER N., PARMAR N., USZKOEIT J., JONES L., GOMEZ A. N., KAISER L. & POLOSHUKIN I. (2017). Attention is all you need. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*. DOI : 10.48550/arXiv.1706.03762.

32] VIG J., GEHRMANN S., BELINKOV Y., QIAN S., NEVO D., SINGER Y. & SHIEBER S. (2020). Investigating gender bias in language models using causal mediation analysis. In H. LAROCHELLE, M. RANZATO, R. HADSELL, M. BALCAN & H. LIN, Éds., *Advances in Neural Information Processing Systems*, volume 33, p. 12388–12401 : Curran Associates, Inc.

33] WALLAT J., BERINGER F., ANAND A. & ANAND A. (2023). Probing bert for ranking abilities. In *Advances in Information Retrieval : 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part II*, p. 255–273, Berlin, Heidelberg : Springer-Verlag. DOI : 10.1007/978-3-031-28238-6_17.

34] ZHAN J., MAO J., LIU Y., ZHANG M. & MA S. (2020). An analysis of bert in document ranking. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, p. 1941–1944, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3397271.3401325.