Recherche Parcimonieuse par Raisonnement Pragmatique

Arthur Satouf[1,2,3,5,6] Gabriel Ben-Zenou[1,4] Benjamin Piwowarski[5,6]
Habiboulaye Amadou-Boubacar[3] Pablo Piantanida[1,2,6]
(1) MILA - Quebec AI Institute and ILLS, Montreal, Canada
(2) Université Paris-Saclay, Gif-sur-Yvette, France
(3) Air Liquide, Paris, France
(4) AMIAD, Palaiseau, France
(5) Sorbonne Université and ISIR, Paris, France
(6) CNRS, Paris, France
arthur.satouf@gmail[.]com, gabriel.ben-zenou@polytechnique[.]edu,
benjamin.piwowarski@cnrs[.]fr,
habiboulaye.amadou-boubacar@airliquide[.]com,
pablo.piantanida@cnrs[.]fr

RÉSUMÉ

Les méthodes actuelles de recherche d'information (RI) neuronales parcimonieuses et, dans une moindre mesure, les modèles plus traditionnels tels que BM25, ne prennent pas en compte les interactions complexes des termes de la représentation d'un même document. Dans cet article, nous montrons comment le cadre linguistique du *Rational Speech Act* (RSA), employé pour minimiser le nombre de caractéristiques à communiquer pour identifier un objet dans un ensemble, peut être adapté au cas de la RI – en particulier face au grand nombre de caractéristiques potentielles (ici, des tokens). Le RSA module dynamiquement les interactions token-document en tenant compte de l'influence des autres documents du corpus, permettant un meilleur contraste entre les représentations de chaque document. Nos expérimentations montrent que l'intégration du RSA améliore de manière systématique plusieurs modèles de RI et atteint des performances à l'état de l'art sur des jeux de données hors distribution du benchmark BEIR.

ABSTRACT

Rational Retrieval Acts : Leveraging Pragmatic Reasoning to Improve Sparse Retrieval

Current sparse neural information retrieval (IR) methods, and to a lesser extent more traditional models such as BM25, do not take into account the complex interplay between different term weights when representing a single document. In this paper, we show how the Rational Speech Acts (RSA), a linguistics framework used to minimize the number of features to be communicated when identifying an object in a set, can be adapted to the IR case – and in particular to the high number of potential features (here, tokens). RSA dynamically modulates token-document interactions by considering the influence of other documents in the dataset, better contrasting document representations. Experiments show that incorporating RSA consistently improves multiple sparse retrieval models and achieves state-of-the-art performance on out-of-domain datasets from the BEIR benchmark.

MOTS-CLÉS : Recherche d'information neuronale, Recherche parcimonieuse, Actes de langage rationnels, Raisonnement pragmatique, Pragmatique linguistique pour la recherche d'information.

KEYWORDS: Neural Information Retrieval, Sparse Retrieval, Rational Speech Acts, Pragmatic

1 Introduction

Le domaine de la RI a évolué rapidement depuis l'adoption de l'architecture *Transformer* pour concevoir divers types de modèles, allant de *cross-encoders* efficaces mais moins performants (Nogueira *et al.*, 2019) à des modèles fondés sur des architectures duales (*bi-encoders*) atteignant de meilleures performances (Hofstätter *et al.*, 2021; Formal *et al.*, 2024). Parmi ces derniers, les modèles neuronaux et parcimonieux de RI – comme SPLADE (Formal *et al.*, 2024) – permettent d'obtenir des résultats supérieurs à ceux obtenus avec des modèles standards – comme BM25 (Robertson & Zaragoza, 2009) – sur la plupart des jeux de données. Toutefois, du côté des représentations de documents, Mackenzie *et al.* (Mackenzie *et al.*, 2021) a montré que les modèles parcimonieux basés sur les Transformers peuvent produire des *wacky weights* (ou poids aberrants), c'est-à-dire des poids correspondant à des termes ayant un pouvoir discriminant limité. Ce problème est aggravé par le nombre généralement réduit de mots-clés dans les requêtes, ainsi que par la faible capacité des modèles existants à contraster deux documents incluant des termes similaires. Même des modèles standards tels que BM25 ne capturent pas l'interaction entre les termes au sein des documents.

Un autre enjeu des représentations parcimonieuses concerne la requête (côté utilisateur). Azzopardi et Zuccon (Azzopardi & Zuccon, 2019) décrivent les utilisateurs comme des agents rationnels lorsqu'ils interagissent avec un moteur de recherche : ils cherchent à réduire le nombre de mots-clés qu'ils saisissent. Ces théories fondées sur l'économie de termes en RI font écho à celle de Grice en linguistique (Grice, 1975), centrée sur la manière dont les personnes véhiculent du sens au-delà de ce qui est explicitement énoncé. Ces idées ont été formalisées mathématiquement dans le cadre du *Rational Speech Act* (RSA), proposé par Frank et Goodman (Frank & Goodman, 2012). Le RSA modélise une communication humaine entre un locuteur et un auditeur, supposés partager un savoir commun, agir en tant qu'agents pragmatiques et souhaiter réduire le coût de communication (nombre de termes énoncés) tout en conservant une signification non ambiguë.

Alors que les utilisateurs se comportent comme des locuteurs pragmatiques (Azzopardi & Zuccon, 2019), les modèles de RI existants ne sont pas des auditeurs pragmatiques (Mackenzie *et al.*, 2021). Nous soutenons ici que cet écart peut être réduit à l'aide du RSA, en particulier pour l'adaptation hors domaine. Plus précisément, nous montrons qu'appliquer le RSA à la RI, afin de transformer des représentations de documents littérales en représentations pragmatiques, améliore les performances des modèles parcimonieux. Les réseaux neuronaux et les modèles modernes de RI capturent déjà du sens au-delà de la sémantique des tokens pris individuellement, par exemple via les mécanismes d'attention ; néanmoins, nous distinguons ici des représentations *littérales* (avant application du RSA) et des représentations *pragmatiques* (après traitement par le RSA).

Le terme *pragmatique* doit être compris ici au sens du concept de comportement pragmatique de Bunt (Bunt, 2017), c'est-à-dire agir en tenant compte du contexte. Le RSA intègre ce contexte en considérant des informations externes – dans notre cas, l'ensemble des documents potentiellement requêtés et l'ensemble des tokens du vocabulaire.

Cohn-Gordon et al. (Cohn-Gordon *et al.*, 2018; Cohn-Gordon & Goodman, 2019) ont été les premiers à appliquer le RSA à la génération de légendes d'images et à la traduction automatique afin de réduire l'ambiguïté. En génération de légendes, le RSA raffine la sortie d'un modèle de génération de langage, en garantissant que la légende produite distingue l'image cible des autres images du

jeu de données. En traduction, le RSA modifie la traduction anglaise d'une phrase allemande afin d'assurer une signification non ambiguë relativement aux autres phrases de l'ensemble à traduire. Dans les deux cas, l'objectif est de produire l'énoncé le plus adapté possible compte tenu du contexte – lequel est l'ensemble des significations possibles à véhiculer – c'est-à-dire, toujours au sens de Bunt (Bunt, 2017), de produire l'énoncé le plus pragmatique. Le RSA rapproche les modèles du niveau de pragmatisme humain. De manière analogue aux tâches de langage, nous montrons dans cet article que le RSA améliore la RI en rendant les représentations documentaires plus pragmatiques, alignant ainsi les modèles sur le comportement d'un utilisateur humain rédigeant une requête.

Le RSA a également été utilisé pour modéliser divers phénomènes linguistiques (Degen, 2023; Goodman & Frank, 2016; Degen *et al.*, 2015; Frank, 2016), et sa simplicité ainsi que son explicabilité en font un outil puissant pour modéliser le comportement communicatif humain. Cependant, cette simplicité pose aussi des défis lorsqu'il s'agit de faire passer le RSA à l'échelle dans des scénarios réels avec des jeux de données nettement plus grands. À ce jour, seuls quelques travaux ont réussi à intégrer RSA aux modèles de langue récents, souvent au prix d'adaptations du mécanisme du RSA lui-même (Cohn-Gordon *et al.*, 2018; Cohn-Gordon & Goodman, 2019; Shen *et al.*, 2019; Kim *et al.*, 2020, 2021). De même, appliquer le RSA à la RI requiert des ajustements pour gérer de grands lexiques, en tirant parti des sorties parcimonieuses des modèles de RI.

Dans cet article, nous apportons les contributions suivantes :

— Nous introduisons le *Rational Retrieval Acts* (RRA), une adaptation du RSA (Frank & Goodman, 2012) aux modèles de RI parcimonieux, en l'étendant soigneusement pour gérer un grand nombre de documents et de tokens.

— Nous obtenons des gains de performance significatifs en RI parcimonieuse hors domaine, sans augmenter les coûts d'inférence.

Le code permettant de reproduire les résultats de cet article est accessible dans son intégralité ici : `https://github.com/arthur-75/Rational-Retrieval-Acts` (Satouf *et al.*, 2025) .

2 Les Rational Retrieval Acts (RRA)

Pour appliquer le RSA à la RI, nous faisons correspondre les tokens du vocabulaire $\mathcal{T}$ aux *énoncés* (*utterance*) du RSA et les documents de l'ensemble $\mathcal{D}$ aux *significations* (*meanings*) du RSA. Par souci de clarté, nous utilisons la terminologie de la RI (tokens, documents) plutôt que celle du RSA (énoncés, significations).

2.1 Intégration du RSA

Le RSA suppose initialement que l'on peut définir un poids positif $\mathcal{L}(t, d) \in \mathbb{R}^+$ pour chaque paire token-document $(t, d) \in \mathcal{T} \times \mathcal{D}$. La fonction $\mathcal{L}$ est appelée *lexique initial*. Dans cet article, nous proposons de dériver $\mathcal{L}(t, d)$ directement à partir des poids $w_{t,d}$ prédits par un modèle de RI clairsemé via l'équation suivante :

$$\mathcal{L}(t, d) = f(w_{t,d}) \,, \tag{1}$$

où f représente une fonction de transformation initiale judicieusement choisie.

Le RSA définit ensuite la distribution de l'auditeur littéral sur les documents conditionnellement à chaque token, $L_0(d|t)$.

Elle représente la manière dont le modèle de RI interprète t, c'est-à-dire la probabilité que chaque document d soit pertinent si t apparaît dans la requête.

La formule définissant L_0 inclut également un terme a priori $\mathbf{P}(d)$, qui peut jouer un rôle analogue à celui des approches probabilistes en RI, par exemple celle menée par Ponte et Croft (Ponte & Croft, 1998). La formulation exacte donnée par le RSA est :

$$L_0(d|t) = \frac{\mathbf{P}(d) \cdot \mathcal{L}(t,d)}{\sum_{d' \in \mathcal{D}} \mathbf{P}(d') \cdot \mathcal{L}(t,d')} , \tag{2}$$

où l'on peut écrire le facteur de normalisation dépendant de t comme $Z_t^{(0)} = \sum_{d' \in \mathcal{D}} \mathbf{P}(d') \cdot \mathcal{L}(t,d')$.

En l'absence de connaissance a priori sur les documents, $\mathbf{P}(d)$ est constant pour tous les documents, et l'auditeur littéral se réduit aux poids normalisés par terme $w_{t,d}$ issus du modèle de recherche.

Le pragmatisme est ensuite introduit via le locuteur pragmatique $S_1(t|d)$, une distribution sur les tokens conditionnée par les documents.

Cette distribution modélise la probabilité qu'un utilisateur pragmatique sélectionne un token pour décrire un document donné, en supposant qu'il interprète les associations token-document de manière similaire au modèle de RI. Autrement dit, les utilisateurs s'alignent sur la représentation apprise par le modèle des relations terme-document, mais affinent leur choix de tokens en fonction de l'ensemble des documents et des tokens. Le RSA introduit un hyperparamètre α pour contrôler le degré de pragmatisme de l'utilisateur. Nous discutons notre choix de α en section 3.1. RSA définit S_1 comme suit :

$$S_1(t|d) = \frac{\exp(\alpha \cdot \log(L_0(d|t)))}{Z_d^{(1)}} , \tag{3}$$

où $Z_d^{(1)}$ est le facteur de normalisation pour un d donné.

Si le locuteur pragmatique est un modèle du comportement de l'utilisateur lors de la formulation d'une requête, ce qui nous intéresse est la manière dont le système de RI doit réagir à ce comportement pragmatique humain. Dans le RSA, cette réaction est modélisée par l'auditeur pragmatique $L_1(d|t)$. Sa définition dans le RSA est la suivante :

$$L_1(d|t) = \frac{\mathbf{P}(d) \cdot S_1(t|d)}{Z_t^{(1)}} , \tag{4}$$

où $Z_t^{(1)}$ est le facteur de normalisation pour un t donné.

On pourrait appliquer le RSA de manière itérative, en calculant une nouvelle paire locuteur-auditeur pragmatiques au-dessus du dernier auditeur pragmatique. Toutefois, les travaux précédents sur le RSA choisissent généralement de ne pas le faire (Cohn-Gordon *et al.*, 2018; Cohn-Gordon & Goodman, 2019), d'autant que le paramètre α permet déjà d'équilibrer le niveau de profondeur pragmatique du RSA.

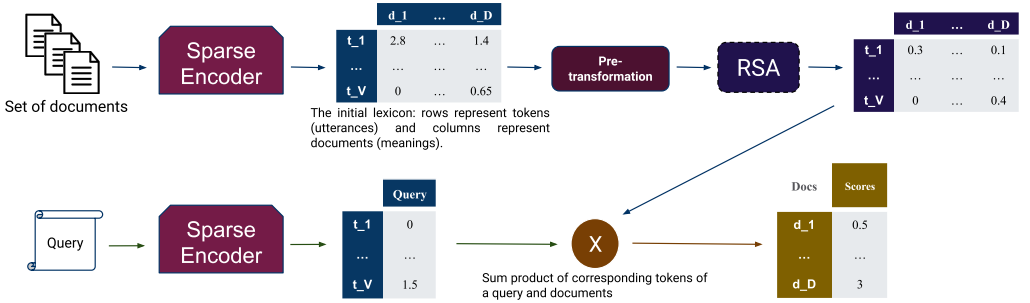

FIGURE 1 – Architecture du Rational Retrieval Act (RRA). La pertinence d'un document $L_1(d|t)$ est calculée par le RSA, en supposant un utilisateur pragmatique (*locuteur*) lors de la formulation de la requête. Le classement final est calculé par produit scalaire entre les représentations de la requête et du document (Eq. (10)).

2.2 Exploiter la parcimonie en RI

Un calcul naïf des distributions du locuteur et de l'auditeur via les équations (3) et (4) n'est pas possible en RI au regard des ressources mémoire, compte tenu du grand nombre de documents et de termes (les matrices S_1 et L_1 ont toutes deux une taille $|\mathcal{T}| \times |\mathcal{D}|$). Pour surmonter ce problème de dimensionnalité, une approche consiste à réduire l'espace des significations possibles en excluant les documents ayant un faible score de similarité avec la requête (Andreas & Klein, 2016; Monroe *et al.*, 2017). Toutefois, une telle stratégie de réduction entre en conflit avec notre besoin d'exécuter un RSA fixe sur l'ensemble de la base de documents en une seule fois. Cohn-Gordon *et al.* (Cohn-Gordon *et al.*, 2018; Cohn-Gordon & Goodman, 2019) procèdent, eux, en exploitant l'aspect séquentiel du texte généré afin de réduire l'espace des énoncés.

De façon similaire, nous proposons une technique qui s'appuie sur la *parcimonie* des modèles de RI parcimonieux auxquels nous appliquons le RSA. Cette parcimonie signifie que la plupart des tokens du vocabulaire n'apparaissent pas dans chaque document : pour la majorité des $(t, d) \in \mathcal{T} \times \mathcal{D}$, $w_{t,d} = 0$. L'application de la fonction de transformation non linéaire f à la sortie des modèles parcimonieux conduit donc la plupart des entrées du lexique RSA $\mathcal{L}(t, d)$ à être égales à $f(0)$. Nous montrons ci-dessous que décomposer $\mathcal{L}(t, d)$ en deux facteurs $l_t^{(0)}$ et $l_d^{(0)}$ correspondant respectivement à l'importance globale d'un terme et d'un document pour toutes les paires (t, d) telles que $w_{t,d} = 0$ permet de réduire drastiquement le stockage mémoire requis, de $|\mathcal{T}| \times |\mathcal{D}|$ à $|\mathcal{T}| + |\mathcal{D}|$, auquel il faut ajouter la taille (faible) du tenseur de poids non nuls produit par le modèle de RI.

Formellement, nous notons d'abord $\mathcal{D}_t$ l'ensemble des documents dans lesquels le token t a un poids non nul ($w_{t,d} \neq 0$) et, de manière analogue, $\mathcal{T}_d$ l'ensemble des tokens ayant un poids non nul pour le document d. Alors, pour tout token t et tout document d tels que $d \notin \mathcal{D}_t$,

$$L_0(d \mid t) = \underbrace{\frac{\mathbf{P}(d) \cdot f(0)}{\sum_{d' \in \mathcal{D}_t} \mathbf{P}(d') \cdot f(w_{t,d'}) + \sum_{d' \notin \mathcal{D}_t} \mathbf{P}(d') \cdot f(0)}}_{l_t^{(0)}} \times \underbrace{1}_{l_d^{(0)}} . \tag{5}$$

En nous appuyant sur cette décomposition initiale, nous pouvons ensuite réécrire les équations (3)

et (4) pour tout token t et tout document d tels que $d \notin \mathcal{D}_t$. Tout d'abord, pour le locuteur pragmatique, nous pouvons réécrire l'équation (3) $\forall t \notin \mathcal{T}_d$:

$$S_1(t \mid d) = \underbrace{\exp\left(\alpha \log l_t^{(0)}\right)}_{s_t^{(1)}} \times \underbrace{\exp\left(\alpha \log l_d^{(0)}\right) / Z_d^{(1)}}_{s_d^{(1)}}, \tag{6}$$

où le facteur de normalisation $Z_d^{(1)}$ est calculé comme

$$Z_d^{(1)} = \sum_{t \in \mathcal{T}_d} \exp\left(\alpha L_0(d|t)\right) + \sum_{t \notin \mathcal{T}_d} \exp\left(\alpha \log(l_t^{(0)} l_d^{(0)})\right). \tag{7}$$

À partir de cette reformulation du locuteur pragmatique, nous pouvons réécrire l'auditeur pragmatique (4) pour tous les $d \notin \mathcal{D}_t$:

$$L_1(d \mid t) = \underbrace{s_t^{(1)} / Z_t^{(1)}}_{l_t^{(1)}} \times \underbrace{s_d^{(1)}}_{l_d^{(1)}}, \tag{8}$$

avec le facteur de normalisation $Z_t^{(1)}$ calculé comme

$$Z_t^{(1)} = \sum_{d' \in \mathcal{D}_t} S_1(d|t) + \sum_{d' \notin \mathcal{D}_t} s_t^{(1)} s_d^{(1)}. \tag{9}$$

Insistons sur le fait que ceci ne modifie en rien le processus du RSA : ce travail de reformulation est effectué pour des raisons de stockage mémoire, et est nécessaire lors de l'application du RRA à des bases de nombreux documents ou à des vocabulaires de grande taille.

2.3 Scorer des documents

La représentation finale d'un document après application du RSA est $(L_1(d|t))_{t \in \mathcal{T}}$. Bien que l'alternance du RSA entre le locuteur et l'auditeur puisse théoriquement être exécutée sur plusieurs itérations, les applications pratiques du RSA à la modélisation du comportement linguistique humain se limitent généralement à une seule itération. Frank (Frank, 2016) montre que α et le nombre d'itérations se compensent en partie. Nous fixons donc le nombre d'itérations à 1 et détaillons notre choix de α en section 3.1.

Enfin, comme illustré en Figure 1, après le RSA, le score d'un document donné est :

$$\text{score}(q, d) = \sum_{t \in \mathcal{T}_d} w_{t,q} \times L_1(d|t) + l_d^{(1)} \times \sum_{t \notin \mathcal{T}_d} w_{t,q} \times l_t^{(1)}, \tag{10}$$

où $w_{t,q}$ est la valeur associée au token t dans la représentation de la requête sur le vocabulaire du modèle de RI.

Au total, les équations (5) à (9) définissent le calcul des représentations de documents dans cette nouvelle version du RSA appliqué au cadre de la RI, le *Rational Retrieval Acts (RRA)*. Par rapport aux équations originales ((2) à (4)), cela permet un calcul efficace en mémoire des locuteurs et auditeurs pragmatiques. Au moment de la recherche, l'équation (10) permet d'exploiter les index inversés $s_t^{(1)}$, $s_d^{(1)}$, $l_d^{(1)}$ et $l_t^{(1)}$ pour une recherche rapide, moyennant des adaptations mineures. RRA peut donc être appliqué à de larges collections de documents.

3 Expérimentations

3.1 Protocole expérimental

Pour nos expériences, nous implémentons le RRA illustré par la Figure 1. Les encodeurs parcimonieux retenus ainsi que les jeux de données de documents et de requêtes sont détaillés dans la section 3.2. Le RSA requiert la définition d'une distribution *a priori* et d'un paramètre α. Pour celle-ci, nous supposons une distribution uniforme sur les documents puisqu'aucune connaissance *a priori* spécifique n'est fournie. Cela se traduit par $\mathbf{P}(d) = \frac{1}{|\mathcal{D}|}$ pour tout $d \in \mathcal{D}$.

Concernant α, lorsque sa valeur tend vers l'infini, l'équation 3 montre clairement que la densité de probabilité définie par le locuteur pragmatique tend vers une distribution de Dirac. À l'inverse, lorsque α tend vers 0, elle tend vers une distribution uniforme. Dans le premier cas, un seul document reçoit une probabilité non nulle (égale à 1), ce qui rend impossible de distinguer les documents suivants ; dans le second cas, tous les documents ont la même importance et ne peuvent donc pas être distingués. Des travaux antérieurs sur le RSA suggèrent que $\alpha = 1$ fournit le meilleur ajustement entre comportement simulé et comportement humain. Cependant, en pratique, le choix de α peut dépendre des caractéristiques du jeu de données et des paramètres du modèle de recherche. La Figure 2 montre le comportement du modèle S-RRA (Splade+RSA) pour différentes valeurs de α : nous observons une différence de performance pouvant atteindre 10 points de nDCG@10 sur le jeu de données TREC-COVID selon le choix du paramètre α. On s'attend donc à ce que les performances soient bornées entre ces deux extrêmes et qu'il existe un α optimal. Les performances restent toutefois assez stables pour le reste des datasets, et nous pouvons prédire une valeur proche de l'optimum en validant α sur quelques requêtes synthétiques pour chaque jeu de données. Nous suivons la méthodologie de Bonifacio et al. (Bonifacio *et al.*, 2022), et échantillonnons 500 documents du jeu de données étudié. Pour chaque document, nous utilisons LLAMA3-8B (Dubey *et al.*, 2024) afin de générer une unique requête pertinente. Plus précisément, nous instruisons le modèle pour qu'il crée un ensemble diversifié de requêtes. Ce processus produit un jeu de données synthétique de RI constitué de paires requête-document considérées comme vraies. Cela nous permet ensuite de sélectionner la valeur de α qui maximise le nDCG@10 sur le jeu synthétique. Notons que ce jeu de données est relativement petit, et que des expériences utilisant ces mêmes requêtes pour affiner un modèle parsemé de RI n'ont donné aucune amélioration de performance.

Enfin, des expériences préliminaires sur le jeu de données MS-Marco ont montré que la fonction de pré-transformation $f : x \to 1 + x$ était la meilleure parmi les transformations suivantes : $f : x \to x$, $\log(1+x)$, $\exp(x)$, $1+x$, $\lambda \cdot x$, et $\tanh(x)$. Nous avons également observé que les fonctions vérifiant $f(0) \neq 0$ donnent de meilleurs résultats que les autres, probablement parce que cela évite de trop accentuer le contraste des poids de documents $L_1(t|d)$.

3.2 Baselines et jeux de données

Compte tenu de ses performances élevées, nous adoptons SPLADE (Formal *et al.*, 2024) (plus précisément `splade-v3` [1]) comme modèle principal de RI. Nous évaluons sur le benchmark BEIR (Thakur *et al.*, 2021) en configuration zero-shot et, après avoir vérifié l'efficacité de SPLADE, appliquons le RRA de la même manière à d'autres modèles parcimonieux (SPARTA (Zhao *et al.*, 2020), BM25,

1. https://huggingface.co/naver/splade-v3

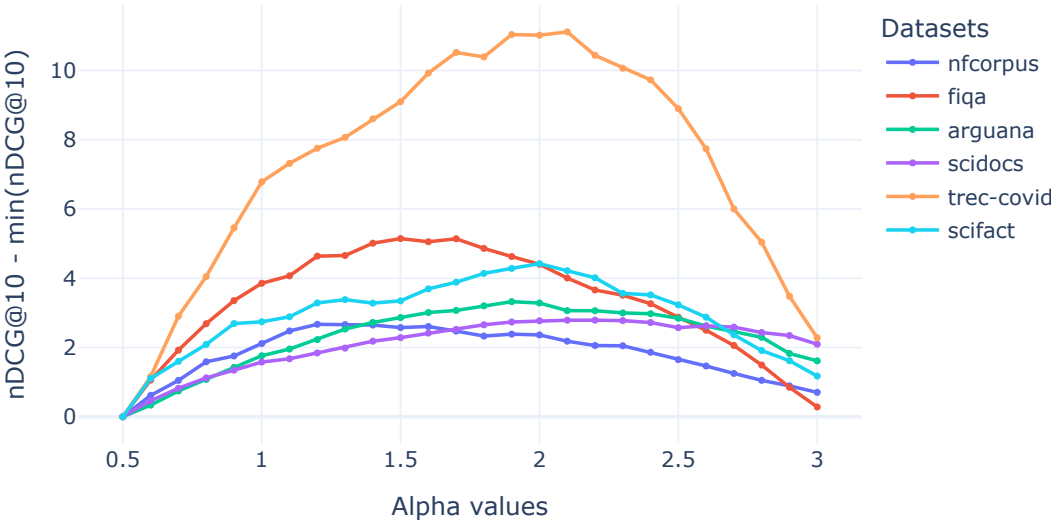

FIGURE 2 – Impact de α sur les performances de S-RRA (nDCG@10). Des valeurs non optimales de α peuvent dégrader significativement les résultats.

TABLE 1 – Tableau 1 : nDCG@10 sur les différents jeux de données pour les modèles parcimonieux et leurs variantes enrichies par RRA (-RRA). En gras : le meilleur score. En souligné : les scores significatifs (p<0,05, test t apparié).

nDCG@10	Splade	S-RRA	Sparta	SP-RRA	bm25	bm-RRA	Unicoil	U-RRA	Deep-Impact	D-RRA
ArguAna	48.9	**50.4**	37.1	**44.1**	34.8	**38.4**	37.4	**41.1**	31.8	**41.4**
FiQA-2018	37.8	**38.3**	19.8	**24.1**	23.8	**23.9**	28.9	**29.5**	27.8	**29.2**
NFCorpus	36.4	**36.6**	30.1	**30.7**	32.5	**32.7**	33.3	**33.6**	27.1	**27.8**
Quora	81.4	**84.0**	62.5	**71.5**	78.9	**79.1**	66.5	**77.2**	66.0	**68.0**
SCIDOCS	15.5	**16.6**	12.6	**14.4**	15.8	**15.9**	14.4	**15.9**	13.4	**14.6**
SciFact	71.6	**73.2**	59.8	**62.4**	66.2	**67.2**	68.6	**69.1**	59.3	**61.9**
TREC-Covid	74.2	**74.7**	53.8	**59.4**	65.3	**66.8**	64.4	**65.7**	56.2	**62.2**
Touch2020	32.2	**32.3**	17.7	**21.6**	36.7	**37.2**	29.1	**30.0**	26.0	**26.8**
CQADupStack	34.5	**35.7**	27.5	**30.4**	30.0	**31.3**	30.2	**31.2**	28.5	**30.4**
Moyenne	48.1	**49.1**	35.7	**39.8**	42.7	**43.6**	41.4	**43.7**	37.3	**40.3**
Gain (points)		**+1**		**+4.2**		**+0.9**		**+2.3**		**+2.9**

Unicoil (Lin & Ma, 2021), DeepImpact (Mallia *et al.*, 2021)). Chaque modèle est comparé à sa version enrichie par RRA (par exemple, SPLADE vs. SPLADE+RSA (S-RRA) et SPARTA vs. SPARTA+RSA (SP-RRA)), comme montré dans le Tableau 1. Pour implémenter ces méthodes, nous nous appuyons sur des outils open-source tels que Anserini, BEIR et les dépôts de code originaux des modèles. Nous rapportons également dans le Tableau 2 les résultats de SparseEmbed (Kong *et al.*, 2023), de ColBERT-v2 (Santhanam *et al.*, 2022), ainsi que les performances de BM25 seul, avec un cross-encodeur, ou utilisé avec une expansion de documents via docT5query(D2Q) (Cheriton, 2019), que nous reprenons de (Wang *et al.*, 2020).

3.3 Résultats

Nos résultats expérimentaux (Tableau 1) mettent en évidence l'efficacité du RRA (colonnes " *-RRA ") pour améliorer les modèles de recherche parcimonieux sur les jeux de données BEIR.

TABLE 2 – Comparaison des scores nDCG@10. Notre modèle S-RRA vs. l'état de l'art. En gras : le meilleur score (score le plus élevé).

nDCG@10	BM25	Col BERTv2	BM25 +CE	TAS-B	D2Q	Sparse Embed	Splade	s-RRA
ArguAna	31.5	46.3	31.1	42.9	34.9	**51.2**	48.9	50.4
CimateFever	21.3	17.6	25.3	22.8	20.1	21.8	25.0	**27.5**
DBPedia	31.3	44.6	40.9	38.4	33.1	**45.7**	44.3	44.4
Fever	75.3	78.5	**81.9**	70.0	71.4	79.6	80.5	81.4
FiQA-2018	23.6	35.6	34.7	30.0	29.1	33.5	37.8	**38.3**
HotpotQA	60.3	66.7	**70.7**	58.4	58.0	69.7	69.5	70.3
NFCorpus	32.5	33.8	35.0	31.9	32.8	34.1	36.4	**36.6**
NQ	32.9	56.2	53.3	46.3	39.9	54.4	58.3	**58.9**
Quora	78.9	85.2	82.5	**83.5**	80.2	84.9	81.4	84.0
SCIDOCS	15.8	15.4	16.6	14.9	16.2	16.0	15.5	**16.6**
SciFact	66.5	69.3	68.8	64.3	67.5	70.6	71.6	**73.2**
TREC-Covid	65.6	73.8	75.7	48.1	71.3	72.4	74.2	**74.7**
Touch2020	36.7	26.3	27.1	16.2	34.7	27.3	32.2	**32.3**
Moyenne	44.0	49.9	49.5	43.7	45.3	50.9	52.0	**53.0**

L'impact est particulièrement marqué pour des jeux de données dont les requêtes sont longues relativement aux documents, tels qu'ArguAna (192,98 mots par requête vs. 166,80 mots par document, ratio = 1,2) et Quora (ratio = 0,83), où le RRA améliore significativement les performances. À l'inverse, pour des jeux comme NFCorpus (ratio = 0,01) et Touché-2020 (ratio = 0,02), où les requêtes sont beaucoup plus courtes, les bénéfices de RSA sont moins prononcés. BM25, qui capture déjà l'interaction globale des termes via l'IDF, bénéficie moins du RRA car son mécanisme de pondération est intrinsèquement conscient du corpus. Contrairement à Splade et Unicoil, Sparta et DeepImpact utilisent un poids uniforme des tokens de requête et obtiennent des gains substantiels avec RRA. En ajustant dynamiquement l'importance des tokens, le RRA compense l'absence de pondération contextuelle des termes de requête, renforçant la capacité de ces modèles à différencier efficacement les documents. Comme montré dans le Tableau 2, l'intégration du RRA à SPLADE (S-RRA) améliore les performances sur la plupart des jeux de données, avec une augmentation moyenne de nDCG@10 d'un point. Nous formulons l'hypothèse que le RRA raffine les interactions token-document, conduisant à de meilleurs contrastes entre documents. Cependant, son impact est moins marqué sur des jeux tels que NFCorpus, DBPedia et Touché-2020, où la modulation des interactions token-document joue un rôle plus limité dans l'efficacité de la recherche.

Étude de cas : SPARTA sur SciFact. Sur le jeu de données SciFact avec le modèle SPARTA, nous observons que sans RRA, le token **"and"** — ainsi que d'autres mots fréquents — peut recevoir des poids disproportionnellement élevés. Par exemple, dans le premier document, **"and"** reçoit un poids de 1,2515, supérieur à celui de tokens plus informatifs tels que **"statistics"** (1,1672) et **"evolution"** (1,0842). Après application du RSA, ce schéma s'inverse : **"and"** est fortement sous-pondéré à 0,00196, tandis que **"statistics"** et **"evolution"** sont scorés à 0,00633 et 0,00467 respectivement, rendant **"and"** environ trois fois moins saillant. Cet exemple illustre comment le RSA supprime, à l'échelle de la collection, les tokens très fréquents et peu informatifs, tout en amplifiant les tokens plus distinctifs et riches en contenu. Cet ajustement dynamique améliore le contraste et le pouvoir

discriminant des représentations parcimonieuses.

4 Discussion

Dans cet article, nous introduisons une adaptation du *Rational Speech Acts* (RSA) à la recherche d'information (RI), que nous appelons *Rational Retrieval Acts* (RRA). Cette méthodologie est compatible avec tout modèle de RI parcimonieux ; ici, nous l'appliquons en particulier à quatre modèles neuronaux parcimonieux à l'état de l'art ainsi qu'à BM25. Nos résultats montrent que le RRA améliore l'efficacité des modèles sans compromettre l'efficience sur des jeux de données hors domaine. Une limite de notre méthodologie est que la phase RSA du RRA doit être réappliquée à l'ensemble de la collection, y compris aux nouveaux documents, à chaque mise à jour du jeu de données. Enfin, nos travaux soulignent également que les représentations de documents, telles que calculées par les modèles neuronaux parcimonieux de RI, peuvent ne pas être optimales au niveau de la collection, et posent la question du calcul – en particulier pour des jeux de données hors domaine – de meilleures représentations dépendantes de la collection.

5 Reconnaissance

Les auteurs remercient Air Liquide pour son soutien financier ainsi que l'ANR – FRANCE (Agence Nationale de la Recherche) pour son soutien financier au projet GUIDANCE n°ANR-23-IAS1-0003.

Références

ANDREAS J. & KLEIN D. (2016). Reasoning about pragmatics with neural listeners and speakers.

AZZOPARDI L. & ZUCCON G. (2019). Building economic models of human computer interaction. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI EA '19, p. 1–4, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3290607.3298809.

BONIFACIO L., ABONIZIO H., FADAEE M. & NOGUEIRA R. (2022). InPars : Unsupervised Dataset Generation for Information Retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, p. 2387–2392 : Association for Computing Machinery. DOI : 10.1145/3477495.3531863.

BUNT H. (2017). Computational pragmatics. *Oxford University Press*.

CHERITON D. R. (2019). From doc2query to doctttttquery. In *From doc2query to docTTTTTquery*.

COHN-GORDON R. & GOODMAN N. (2019). Lost in machine translation : A method to reduce meaning loss.

COHN-GORDON R., GOODMAN N. & POTTS C. (2018). Pragmatically informative image captioning with character-level inference. *arXiv preprint arXiv :1804.05417*.

DEGEN J. (2023). The rational speech act framework. *Annu. Rev. Linguist.*, **9**(1), 519–540.

DEGEN J., TESSLER M. H. & GOODMAN N. D. (2015). *Wonky worlds : Listeners revise world knowledge when utterances are odd*. Annual Reviews.

DUBEY A., JAUHRI A., PANDEY A., KADIAN A., AND STEPHANE COLLOT A. A., GURURANGAN S., NAYANI R., MITRA R., LI R., HOGAN R., WEN R. B. Z., YANG Z. & ZHAO Z. (2024). The llama 3 herd of models.

FORMAL T., LASSANCE C., PIWOWARSKI B. & CLINCHANT S. (2024). Towards Effective and Efficient Sparse Neural Information Retrieval. *ACM Transactions on Information Systems*, p. 3634912. DOI : 10.1145/3634912.

FRANK M. C. (2016). Rational speech act models of pragmatic reasoning in reference games. *Trends Cogn. Sci.*

FRANK M. C. & GOODMAN N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, **336**, 998 – 998.

GOODMAN N. D. & FRANK M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends Cogn. Sci.*, **20**(11), 818–829.

GRICE H. P. (1975). Logic and conversation.

HOFSTÄTTER S., LIN S.-C., YANG J.-H., LIN J. & HANBURY A. (2021). Efficiently teaching an effective dense retriever with balanced topic aware sampling.

KIM H., KIM B. & KIM G. (2020). Will I Sound Like Me ? Improving Persona Consistency in Dialogues through Pragmatic Self-Consciousness. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 904–916, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.65.

KIM H., KIM B. & KIM G. (2021). Perspective-taking and Pragmatics for Generating Empathetic Responses Focused on Emotion Causes. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 2227–2240, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.170.

KONG W., DUDEK J. M., LI C., ZHANG M. & BENDERSKY M. (2023). Sparseembed : Learning sparse lexical representations with contextual embeddings for retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23)*.

LIN J. & MA X. (2021). A few brief notes on deepimpact, coil, and a conceptual framework for information retrieval techniques.

MACKENZIE J., TROTMAN A. & LIN J. (2021). Wacky Weights in Learned Sparse Representations and the Revenge of Score-at-a-Time Query Evaluation. *ACM Transactions on Information Systems*. DOI : 10.48550/arXiv.2110.11540.

MALLIA A., KHATTAB O., TONELLOTTO N. & SUEL T. (2021). Learning passage impacts for inverted indexes.

MONROE W., HAWKINS R. X., GOODMAN N. D. & POTTS C. (2017). Colors in Context : A Pragmatic Neural Model for Grounded Language Understanding. *Transactions of the Association for Computational Linguistics*, **5**, 325–338. DOI : 10.1162/tacl_a_00064.

NOGUEIRA R., YANG W., CHO K. & LIN J. (2019). Multi-stage document ranking with bert.

PONTE J. & CROFT W. (1998). A language modeling approach to information retrieval. In *ACM SIGIR*.

ROBERTSON S. & ZARAGOZA H. (2009). The probabilistic relevance framework : Bm25 and beyond. *Found. Trends Inf. Retr.*, **3**(4), 333–389. DOI : 10.1561/1500000019.

SANTHANAM K., KHATTAB O., SAAD-FALCON J., POTTS C. & ZAHARIA M. (2022). Colbertv2 : Effective and efficient retrieval via lightweight late interaction.

SATOUF A., ZENOU G. B., PIWOWARSKI B., BOUBACAR H. A. & PIANTANIDA P. (2025). Rational retrieval acts : Leveraging pragmatic reasoning to improve sparse retrieval.

SHEN S., FRIED D., ANDREAS J. & KLEIN D. (2019). Pragmatically Informative Text Generation. DOI : 10.48550/ARXIV.1904.01301.

THAKUR N., REIMERS N., RÜCKLÉ A., SRIVASTAVA A. & GUREVYCH I. (2021). Beir : A heterogenous benchmark for zero-shot evaluation of information retrieval models.

WANG W., WEI F., DONG L., BAO H., YANG N. & ZHOU M. (2020). Minilm : Deep self-attention distillation for task-agnostic compression of pre-trained transformers.

ZHAO T., LU X. & LEE K. (2020). Sparta : Efficient open-domain question answering via sparse transformer matching retrieval.