



# Modèles de langue à la *BERT* et documents longs

Eric Gaussier

*With J. Chagnon, Y. Cinar, Q. Grail, M. Li, J. Perez, D. Popa*

MIAI - LIG - UGA

5 octobre 2021

# Table of Contents

## **Introduction**

*G-BERT : globalizing BERT*

Sélection de blocs pour la RI

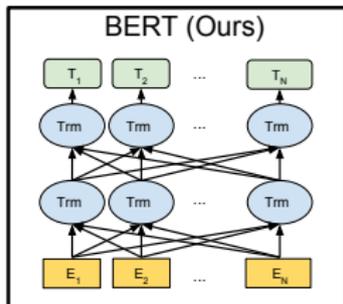
Discussion

# Modèles de langue, TAL et RI

## Les modèles de langue actuels

- ▶ Modèles de langue présents dans de nombreuses applications : traduction automatique, génération, dialogue, questions-réponses, RI *ad hoc*, ...
- ▶ Modèles pré-entraînés avec une supervision naturelle (*self-supervised*)
- ▶ Fournissent des plongements lexicaux (*word embeddings*) contextualisés
- ▶ BERT en est un exemple prototypique

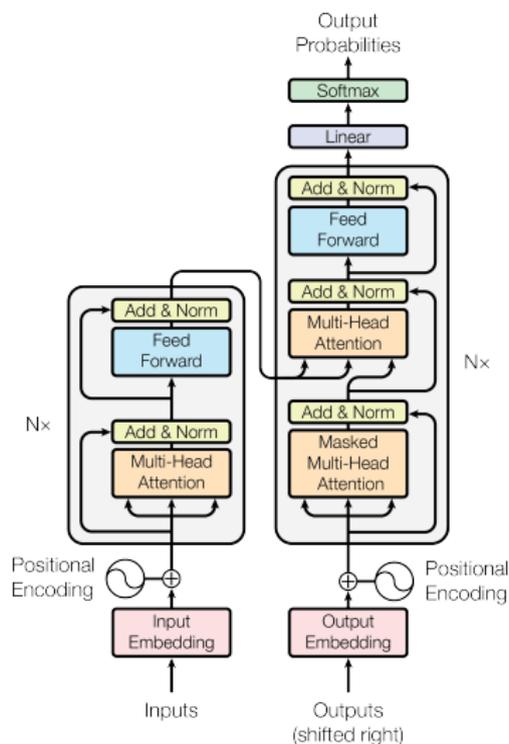
## Les modèles de langue à la BERT



- ▶ Modèles de langue bidirectionnels aux performances "état-de-l'art"
- ▶ Pré-entraînement puis adaptation aux collections ciblées
- ▶ Disponible en plusieurs langues et sur plusieurs domaines (camembert, flaubert)
- ▶ Brique fondamentale : transformeur

*Image tirée de Devlin et al.*

# Les transformeurs (image tirée de Vaswani et al.)



# Propriétés des transformeurs (1)

*Attention is all you need!*

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

# Propriétés des transformeurs (1)

*Attention is all you need!*

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

## Propriétés (biais inductifs)

- ▶ Modélisation directe les dépendances locales et à longue distance
- ▶ Les têtes multiples (8 par ex.) permettent de capturer différents types de dépendances
- ▶ Les couches successives (6 par ex.) permettent de capturer des dépendances complexes ( $\Rightarrow$  structure)

# Interlude sur les biais inductifs

## Pas d'apprentissage sans biais

- ▶ Biais lié :
  - ▶ Aux propriétés du modèle (par ex. CNN)
  - ▶ Aux fonctions objectifs (stratégie pour éviter les répétitions et les hallucinations en génération par ex.)
  - ▶ Aux données et aux connaissances (modèles flexibles permettant de les prendre en compte)
- ▶ Biais permet d'apprendre un modèle qui "généralise bien"
- ▶ *La capacité de généralisation des modèles neuronaux actuels pour le TAL reste faible*

Jeux de données SCAN (Lake & Baroni) et COGS (Kim & Linzen)  
- *compositional generalization*

## Propriétés des transformeurs (2)

*Attention is all you need!*

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

### Considérations informatiques

- ▶ Calcul du gradient simple et facilement parallélisable
- ▶ Complexité en  $\mathcal{O}(n^2)$  rend difficile l'utilisation sur des documents longs (512 *tokens*)

# Transformeurs/BERT et documents longs (1)

## Idée générale : travailler sur des blocs de taille raisonnable

1. Limiter la taille des documents de façon arbitraire ou *par une sélection adéquate* - MultiHop QA (Yang *et al.*, Fang *et al.*), RI (à venir)
2. Utiliser une fenêtre glissante (Joshi *et al.* pour la résolution de coréférence)
3. *Utiliser des modèles hiérarchiques* qui permettent de partager l'information entre différents blocs (Tu *et al.*)

## Transformeurs/BERT et documents longs (2)

La bonne approche dépend des caractéristiques requises :  
représentation/compréhension globale ou partielle d'un document,  
positions relatives des informations pertinentes, ...

- ▶ Applications TAL requièrent souvent une représentation fine et complète d'un document (résumé automatique par ex.)
- ▶ La RI *ad hoc* ne requiert qu'une représentation partielle d'un document (représentation fine des seules parties pertinentes)

Nous privilégions des pistes différentes suivant l'application visée

# Table of Contents

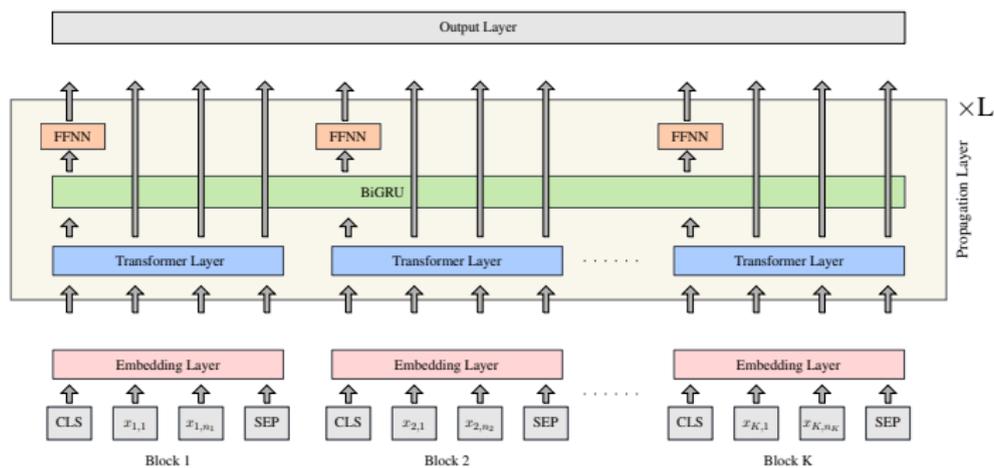
Introduction

***G-BERT : globalizing BERT***

Sélection de blocs pour la RI

Discussion

# BERT global (Grail *et al.*)



## Couche de propagation

La propagation de l'information se fait au sein de chaque couche de transformeurs

- ▶ Chaque document est divisé en  $K$  blocs de taille  $n_k$
- ▶ Plongement (*embedding*) identique à celui de BERT
- ▶ Soient  $U_k^\ell$  la représentation du  $k^{\text{ème}}$  bloc après la couche  $\ell - 1$  et  $T^\ell$  la fonction correspondant au transformeur pré-entraîné de la couche  $\ell$ .  $U_k^{\ell+1}$  est défini par :

$$V_k^\ell = T^\ell(U_k^\ell)$$

$$W_k^\ell = \text{FFNN}(\text{BiGRU}([V_{1,0}^\ell; \dots; V_{K,0}^\ell])_k)$$

$$U_k^{\ell+1} = [W_k^\ell; V_{k,1}^\ell; \dots; V_{k,n_k+1}^\ell]$$

## Application au résumé extractif

- ▶ Entrée : une phrase ; sortie : probabilité que la phrase soit choisie

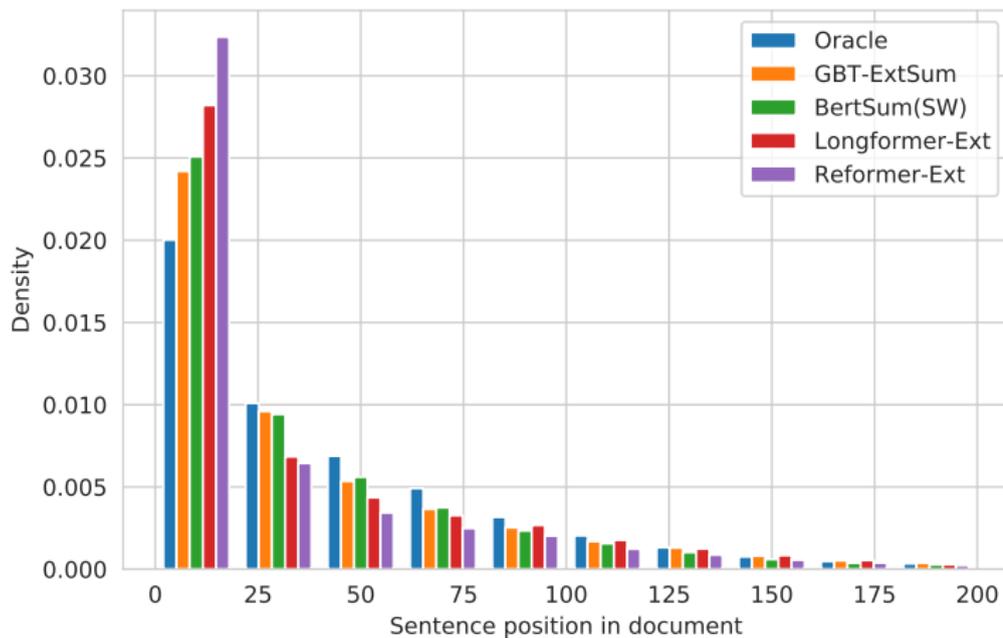
$$y_k = \text{Softmax}(\text{FFNN}(W_K^{L+1})), L = 12$$

- ▶ Version de BERT utilisée : bert-base-uncase
- ▶ Collections de documents longs : ArXiv et Pubmed (Cohan *et al.*)
  - ▶ ArXiv : 203 037/6 436/6 440 docs ; taille moyenne 5038 mots
  - ▶ Pubmed : 119 924/6 633/6 658 docs ; taille moyenne 3235 mots
- ▶ Evaluation : ROUGE ; annotation à la Kedzie (Kedzie *et al.*)
- ▶ Comparaison avec 17 méthodes pour le résumé abstraktif et extractif

# Résultats quantitatifs

Summarizer	PubMed				arXiv				
	RG-1	RG-2	RG-3	RG-L	RG-1	RG-2	RG-3	RG-L	
Oracle	58.15	34.16	24.11	52.99	57.78	30.43	18.41	51.24	
Lead	37.77	13.35	7.64	34.31	35.54	9.50	3.33	31.19	
Abstractive or Mix	Attn-Seq2Seq (Nallapati et al., 2016)	31.55	8.52	7.05	27.38	29.30	6.00	1.77	25.56
	Pntr-Gen-Seq2Seq (See et al.)	35.86	10.22	7.60	29.69	32.06	9.04	2.15	25.16
	Discourse summarizer (Cohan et al., 2018)	38.93	15.37	9.97	35.21	35.80	11.05	3.62	31.80
	TLM-I+E (G,M) (Subramanian et al., 2019)	42.13	16.27	8.82	39.21	41.62	14.69	6.16	38.03
	DANCER PEGASUS (Gidiotis and Tsoumakas, 2020)	46.34	19.97	-	42.42	45.01	17.60	-	40.56
	PEGASUS (Zhang et al., 2019a)	45.97	20.15	-	28.25	44.21	16.95	-	25.67
	BIGBIRD-Pegasus (Zaheer et al., 2020)	46.32	20.65	-	42.33	46.63	19.02	-	41.77
Extractive	SumBasic (Vanderwende et al., 2007)	37.15	11.36	5.42	33.43	29.47	6.95	2.36	26.30
	LexRank (Erkan and Radev, 2004)	39.19	13.89	7.27	34.59	33.85	10.73	4.54	28.99
	LSA (Steinberger and Jezek, 2004)	33.89	9.93	5.04	29.70	29.91	7.42	3.12	25.67
	Sent-CLF (Subramanian et al., 2019)	45.01	19.91	<b>12.13</b>	41.16	34.01	8.71	2.99	30.41
	Sent-PTR (Subramanian et al., 2019)	43.30	17.92	10.67	39.47	42.32	15.63	7.49	38.06
	Bert Ranker (Nogueira and Cho, 2019)	43.67	18.00	10.74	39.22	41.65	13.88	5.92	36.40
	BERTSUMEXT (Liu and Lapata, 2019b)	41.09	15.51	8.64	36.85	41.24	13.01	5.26	36.10
	BERTSUMEXT (SW) (Liu and Lapata, 2019b)	45.01	20.00	12.05	40.43	42.93	15.08	6.01	37.22
	Longformer-Ext (Beltagy et al., 2020)	43.75	17.37	10.18	39.71	45.24	16.88	8.06	40.03
	Reformer-Ext (Kitaev et al., 2020)	42.32	15.91	9.02	38.26	43.26	14.86	6.66	38.10
	GBT-EXTSUM (Ours)	<b>46.87</b>	<b>20.19</b>	12.11	<b>42.68</b>	<b>48.08</b>	<b>19.21</b>	<b>9.58</b>	<b>42.68</b>

# Position des phrases extraites



# Illustration

GOLD	<p>purpose : to investigate whether the glc3a locus harboring the cyp1b1 gene is associated with normal tension glaucoma ( ntg ) in japanese patients.materials and methods : one hundred forty two japanese patients with ntg and 101 japanese healthy controls were recruited . patients exhibiting a comparatively early onset were selected as this suggests that genetic factors may show stronger involvement . genotyping and assessment of allelic diversity was performed on 13 highly polymorphic microsatellite markers in and around the glc3a locus.results:there were decreased frequencies of the 444 allele of d2s0416i and the 258 allele of d2s0425i in cases compared to controls ( <math>p = 0.022</math> and <math>p = 0.034</math> , respectively ) . however , this statistical significance disappeared when corrected ( <math>pc &gt; 0.05</math> ) . we did not find any significant association between the remaining 11 microsatellite markers , including d2s177 , which may be associated with cyp1b1 , and ntg ( <math>p &gt; 0.05</math> ) . conclusions : our study showed no association between the glc3a locus and ntg , suggesting that the cyp1b1 gene , which is reportedly involved in a range of glaucoma phenotypes , may not be an associated factor in the pathogenesis of ntg .</p>
GRT-EXTSUM	<p>1- primary open angle glaucoma ( poag ) is the most common type of glaucoma .</p> <p>15- we excluded individuals who were diagnosed under 20 or over 60 years of age and who had 8.0 d or higher myopic refractive error of spherical equivalence .</p> <p>17- the cases exhibiting a comparatively early onset were selected as they suggest that genetic factors may show stronger involvement . during diagnosis .</p> <p>30- the probability of association was corrected by the bonferroni inequality method , ie , by multiplying the obtained p values with the number of alleles compared .</p> <p>63- only two adjacent markers , d2s0416i and d2s0425i , were significantly positive , as shown in table 2 , and the frequency of the 444 allele of d2s0416i and the 258 allele of d2s0425i were decreased in cases compared to controls ( <math>p = 0.022</math> , or = 0.59 and <math>p = 0.034</math> , or = 0.42 , respectively ) .</p> <p>66- the purpose of this study was to investigate whether the glc3a locus is associated with ntg in japanese subjects , based on results from recent studies reporting that the cyp1b1 gene , located at the glc3a locus on chromosome 2p21 , could be a causative gene in poag as well as pcg . to this end , we genotyped 13 microsatellite markers in and around the glc3a locus . here</p>
BERTSUMEXT (SW)	<p>1- primary open angle glaucoma ( poag ) is the most common type of glaucoma .</p> <p>2- normal tension glaucoma ( ntg ) is an important subset of poag ; while many poag patients have high iop,1 patients with ntg have statistically normal iop.24 the prevalence of the japanese population than among caucasians , and recent studies reported that 92% of poag patients in japan had ntg.58 the diagnosis of glaucoma is based on a combination of factors including optic nerve damage and specific field defects for which iop is the only treatable risk factor .</p> <p>7- of these subjects , 142 were diagnosed with ntg , and 101 were control subjects .</p> <p>20- genomic dna was extracted using the qiamp dna blood mini kit ( qiagen , hilden , germany ) or the guanidine method . in this association study , we selected 13 highly polymorphic microsatellite markers that are located in and around the glc3a locus as shown in figure 1 .</p> <p>28- the number of microsatellite repeats was estimated automatically using the genescan 672 software ( applied biosystems ) by the local southern method with a size marker of gs500 tamra ( applied biosystems ) .</p> <p>22- polymerase chain reaction ( pcr ) was performed in a reaction mixture with a total volume of 12.5 l containing pcr buffer , genomic dna , 0.2 mm dinucleotide triphosphates ( dntps ) , 0.5 m primers , and 0.35 u taq polymerase .</p>

## Conclusion (1)

- ▶ Difficulté de traiter des documents longs avec les modèles de langue actuels (BERT)
- ▶ Il est possible d'étendre (hiérarchiquement) ces modèles pour :
  - ▶ Tenir compte des dépendances globales
  - ▶ Obtenir une représentation complète du document (chaque phrase dépend des autres phrases pour le résumé extractif)
- ▶ Approche adaptée au TAL
- ▶ RI *ad hoc* s'appuie sur des informations de pertinence locale  
→ Comment sélectionner ces informations ?

# Table of Contents

Introduction

*G-BERT : globalizing BERT*

**Sélection de blocs pour la RI**

Discussion

## Analyse de quelques collections standard

Collections retenues : Robust04, Gov2, MQ2007, MQ2008

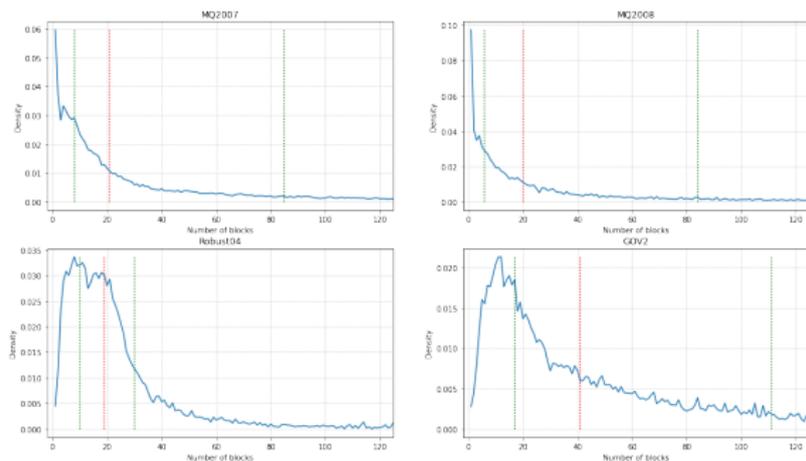
Dataset	MQ2007	MQ2008	GOV2	Robust04
Nb of queries	1,692	784	150	250
<b>Original</b>				
Nb of documents	65,302	14,381	ca. 25M	ca. 0.5M
Nb of unique labeled documents	65,302	14,381	128,010	174,787
Nb of labeled document-query pairs	69,599	15,208	135,352	311,410
<b>BM25 filtered</b>				
Nb of unique documents	-	-	29,769	42,156
Nb of unique labeled documents	-	-	24,682	38,905
Nb of labeled document-query pairs	-	-	26,155	95,336
Share of irrelevant pairs	0.74	0.81	0.80	0.94
Share of relevant pairs	0.20	0.13	0.17	0.05
Share of very relevant pair	0.06	0.6	0.03	<0.01

Table 1. Statistics of the datasets used.

1. Longueur des documents ?
2. Position de l'information pertinente ?
3. Appariement exact ou flou (*exact vs fuzzy matching*) ?

## Longueur des documents ?

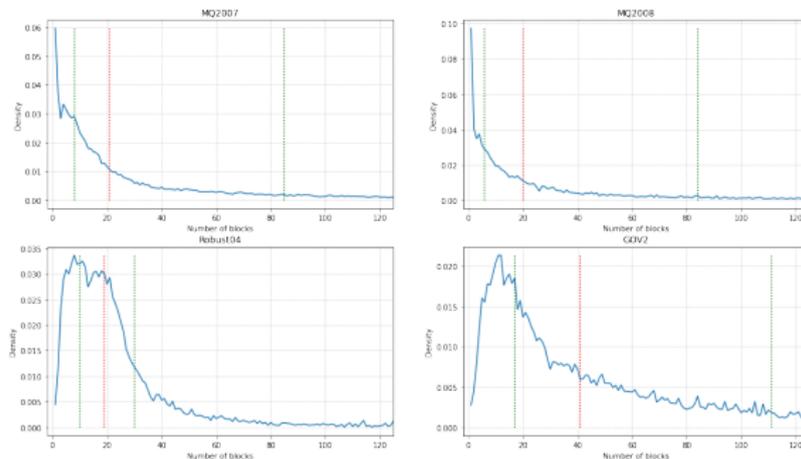
Découpage en blocs suivant la stratégie de CogLTX (Ding *et al.*) -  
1 bloc  $\approx$  62 tokens



*Distribution des documents suivant le nombre de blocs (25-50-75 percentile)*

# Longueur des documents ?

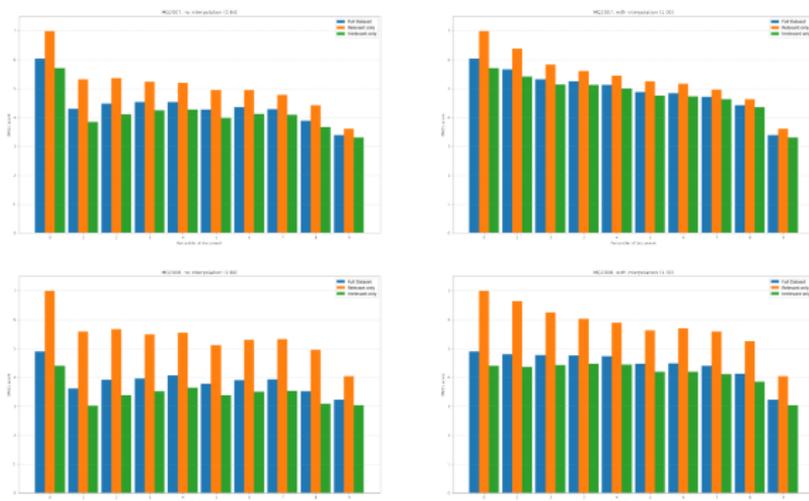
Découpage en blocs suivant la stratégie de CogLTX (Ding *et al.*) -  
1 bloc  $\approx$  62 tokens



Longueur supérieure aux limites des modèles de langue

# Où se trouve l'information pertinente ?

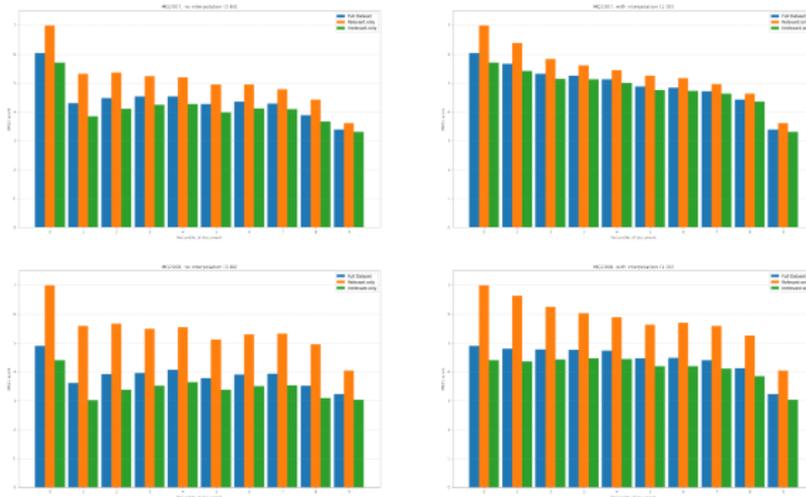
Pertinence bloc  $b$  pour requête  $q \approx \text{RSV}_{\text{BM25}}(q, b)$



*Histogramme pertinence-position pour tous les docs (bleu), les docs pertinents (orange) et les docs non pertinents (vert)*

# Où se trouve l'information pertinente ?

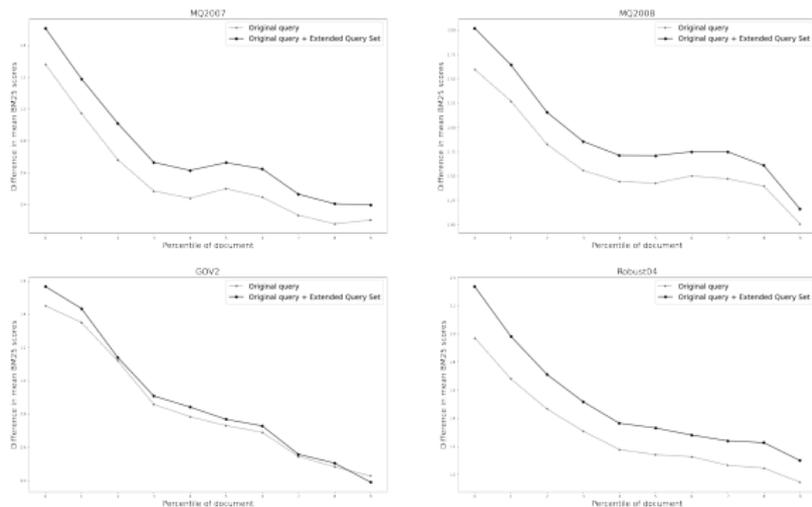
Pertinence bloc  $b$  pour requête  $q \approx \text{RSV}_{\text{BM25}}(q, b)$



L'information pertinente n'est pas concentrée en de rares endroits -  
*verbosity hypothesis*

# Appariement exact ou flou ?

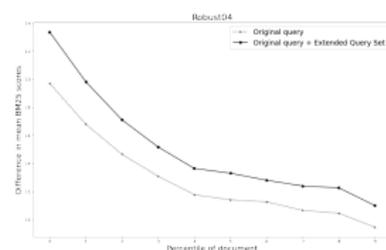
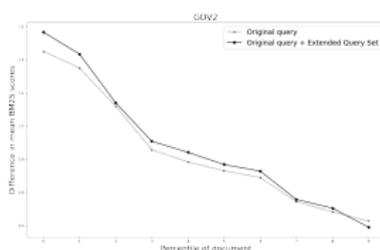
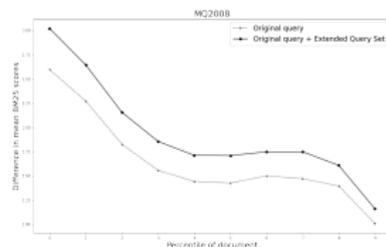
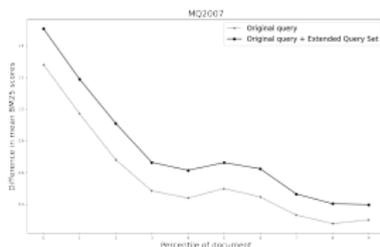
## Comparaison requête originale et requête étendue avec synonymes (WordNet)



*Différence ( $RSV(pert) - RSV(non-pert)$ ) en fonction de la position pour les requêtes originales et les requêtes étendues*

# Appariement exact ou flou ?

Comparaison requête originale et requête étendue avec synonymes (WordNet)

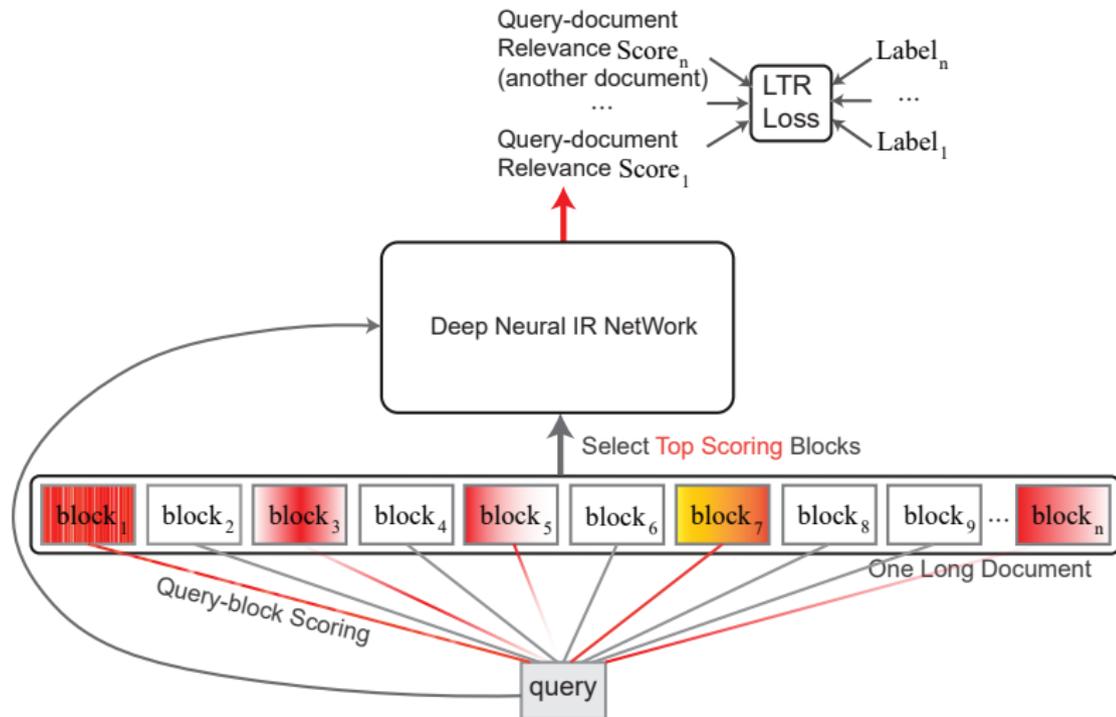


Gains potentiels avec appariement flou

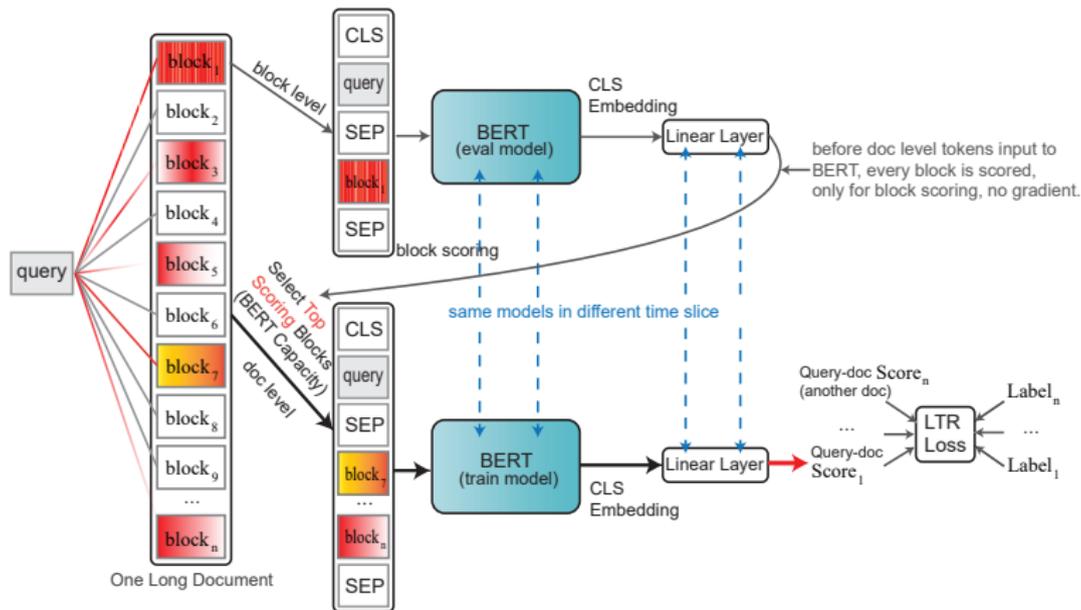
## Stratégies de sélection

- ▶ Sélectionner les blocs n'importe où dans le document en fonction de leur pertinence potentielle
- ▶ Pertinence potentielle :
  - ▶ Score standard RSV - *TF-IDF*, *BM25* ; appariement exact
  - ▶ Score appris en même temps que le modèle de RI ; appariement flou

## Sélection *standard*



## Sélection apprise



## Cadre expérimental

- ▶ Mêmes collections que pour l'analyse + TREC DL 2019 (Craswell *et al.*)
- ▶ Deux modèles privilégiés : Vanilla BERT et PARADE (Li *et al.* 2020)
- ▶ Comparaison avec différentes approches

## Résultats - MQ2008

Model	P@1	P@5	P@10	MAP	nDCG@1	nDCG@5	nDCG@10	nDCG
BM25	0.3839 <sup>†</sup>	0.3252 <sup>†</sup>	0.2384 <sup>†</sup>	0.4486 <sup>†</sup>	0.3316 <sup>†</sup>	0.4303 <sup>†</sup>	0.4799 <sup>†</sup>	0.5225 <sup>†</sup>
DeepRank	0.3992 <sup>†</sup>	0.2816 <sup>†</sup>	0.1920 <sup>†</sup>	0.4356 <sup>†</sup>	0.3641 <sup>†</sup>	0.4373 <sup>†</sup>	0.4672 <sup>†</sup>	0.4917 <sup>†</sup>
DeepRank*	0.482	0.359	0.252	0.498	0.406	0.496	-	-
Vanilla BERT	0.5025	0.3660 <sup>†</sup>	0.2556 <sup>†</sup>	0.5258 <sup>†</sup>	0.4597	0.5165 <sup>†</sup>	0.5542	0.5849
PARADE	0.5101	0.3775 <sup>†</sup>	0.2589	0.5298 <sup>†</sup>	0.4508	0.5236 <sup>†</sup>	0.5580	0.5853
CEDR_KNRM	0.5050	0.3678 <sup>†</sup>	0.2561 <sup>†</sup>	0.5220 <sup>†</sup>	0.4515	0.5151 <sup>†</sup>	0.5488 <sup>†</sup>	0.5794
Random Select	0.5153	0.3763 <sup>†</sup>	0.2551 <sup>†</sup>	0.5141 <sup>†</sup>	0.4438	0.5068 <sup>†</sup>	0.5445 <sup>†</sup>	0.5748 <sup>†</sup>
KeyB(vBERT) <sub>TF-IDF</sub>	0.5166	<b>0.3862</b>	0.2597	0.5318 <sup>†</sup>	0.4649	0.5330	0.5596	0.5869
KeyB(vBERT) <sub>BM25</sub>	0.5165	0.3760 <sup>†</sup>	0.2579 <sup>†</sup>	0.5350	0.4629	0.5317	0.5609	<b>0.5891</b>
KeyB(PARADE5) <sub>BM25</sub>	0.5204	0.3768	0.2589	0.5358	0.4649	0.5365	<b>0.5616</b>	0.5880
KeyB(vBERT) <sub>BinB</sub>	<b>0.5254</b>	0.3819	<b>0.2624</b>	<b>0.5425</b>	<b>0.4661</b>	<b>0.5382</b>	<b>0.5616</b>	<b>0.5891</b>

## Résultats - Gov2

Model	P@1	P@5	P@10	P@20	MAP
BM25	0.6510	0.6054 <sup>†</sup>	0.5792 <sup>†</sup>	0.5362 <sup>†</sup>	0.2331 <sup>†</sup>
DeepRank	0.6453	0.5682 <sup>†</sup>	0.5143 <sup>†</sup>	0.4880 <sup>†</sup>	0.2151 <sup>†</sup>
Vanilla BERT	0.6248 <sup>†</sup>	0.6004 <sup>†</sup>	0.5666 <sup>†</sup>	0.5483 <sup>†</sup>	0.2314 <sup>†</sup>
PARADE	0.6506	0.6535	0.6181 <sup>†</sup>	0.5840 <sup>†</sup>	0.2547 <sup>†</sup>
CEDR_KNRM	0.6913	0.6096 <sup>†</sup>	0.5746 <sup>†</sup>	0.5437 <sup>†</sup>	0.2343 <sup>†</sup>
Random Select	0.6303 <sup>†</sup>	0.6241 <sup>†</sup>	0.5842 <sup>†</sup>	0.5586 <sup>†</sup>	0.2386 <sup>†</sup>
KeyB(vBERT) <sub>TF-IDF</sub>	0.6510 <sup>†</sup>	0.6594	0.6242 <sup>†</sup>	0.5912 <sup>†</sup>	0.2581 <sup>†</sup>
KeyB(vBERT) <sub>BM25</sub>	<b>0.7310</b>	0.6388 <sup>†</sup>	0.6262	0.5795	0.2580 <sup>†</sup>
KeyB(PARADE5) <sub>BM25</sub>	0.7120	<b>0.6818</b>	0.6506	0.6151	<b>0.2715</b>
KeyB(vBERT) <sub>BinB</sub>	0.7113	0.6629	<b>0.6580</b>	<b>0.6179</b>	0.2604 <sup>†</sup>
Model	nDCG@1	nDCG@5	nDCG@10	nDCG@20	nDCG
BM25	0.5034	0.4904 <sup>†</sup>	0.4867 <sup>†</sup>	0.4774 <sup>†</sup>	0.4296 <sup>†</sup>
DeepRank	0.4738 <sup>†</sup>	0.4363 <sup>†</sup>	0.4194 <sup>†</sup>	0.4170 <sup>†</sup>	0.4120 <sup>†</sup>
Vanilla BERT	0.4528 <sup>†</sup>	0.4837 <sup>†</sup>	0.4714 <sup>†</sup>	0.4670 <sup>†</sup>	0.4286 <sup>†</sup>
PARADE	0.5361	0.4831 <sup>†</sup>	0.5068 <sup>†</sup>	0.5015 <sup>†</sup>	0.4412 <sup>†</sup>
CEDR_KNRM	0.5031	0.4581 <sup>†</sup>	0.4618 <sup>†</sup>	0.4626 <sup>†</sup>	0.4274 <sup>†</sup>
Random Select	0.4592 <sup>†</sup>	0.4589 <sup>†</sup>	0.4693 <sup>†</sup>	0.4761 <sup>†</sup>	0.4266 <sup>†</sup>
KeyB(vBERT) <sub>TF-IDF</sub>	0.4897	0.5023 <sup>†</sup>	0.5098 <sup>†</sup>	0.5012 <sup>†</sup>	0.4439 <sup>†</sup>
KeyB(vBERT) <sub>BM25</sub>	0.5467	0.5091 <sup>†</sup>	0.5122	0.5072 <sup>†</sup>	0.4415 <sup>†</sup>
KeyB(PARADE5) <sub>BM25</sub>	<b>0.5572</b>	<b>0.5517</b>	<b>0.5436</b>	<b>0.5390</b>	<b>0.4514</b>
KeyB(vBERT) <sub>BinB</sub>	0.4897	0.4975 <sup>†</sup>	0.5035 <sup>†</sup>	0.5363	0.4414 <sup>†</sup>

## Résultats - TREC DL 2019

Protocole identique à Craswell *et al.*

Model	nDCG@10	MAP
BM25	0.488	0.234
CO-PACRR	0.550	0.231
TK	0.594	0.252
TKL	0.644	0.277
RoBERTa (FirstP)	0.588	0.233
RoBERTa (MaxP)	0.630	0.246
Sparse-Transformer	0.634	0.257
Longformer-QA	0.627	0.255
Transformer-XH	0.646	0.256
QDS-Transformer	0.667	0.278
KeyB(vBERT) <sub>BinB</sub>	<b>0.685</b>	<b>0.283</b>

## Conclusion (2)

- ▶ Approche simple qui permet de traiter de façon adaptée les documents longs
- ▶ Résultats compétitifs, avec moins d'information pour des modèles comme PARADE
- ▶ Stratégie en deux étapes (standard RI neuronale) :
  1. Sélection des  $N$  documents les plus *prometteurs* avec un modèle de RI standard (BM25)
  2. Tri de ces documents avec un modèle de RI neuronal
- ▶ Coût reste important

Travail partiellement publié (Li & Gaussier 2021)

# Table of Contents

Introduction

*G-BERT : globalizing BERT*

Sélection de blocs pour la RI

**Discussion**

## Quelques réflexions/questions

- ▶ **RI neuronale**
  - ▶ Peut-on se passer de la première étape ?
  - ▶ Les modèles à *la BERT* ont-ils les bonnes propriétés (contraintes de la RI) ? En particulier, la contrainte de normalisation par la longueur n'est pas toujours compatible avec la sélection de blocs.

## Quelques réflexions/questions

### ▶ RI neuronale

- ▶ Peut-on se passer de la première étape ?
- ▶ Les modèles à la *BERT* ont-ils les bonnes propriétés (contraintes de la RI) ? En particulier, la contrainte de normalisation par la longueur n'est pas toujours compatible avec la sélection de blocs.

### ▶ TAL et RI

- ▶ Quelle est la capacité de généralisation de ces modèles ? Quels sont les biais inductifs que nous devons prendre en compte ?
- ▶ Leur coût est-il acceptable ? Quelles pistes pour le réduire (distillation, modèles parcimonieux) ?

WALK LEFT AND RUN --- JUMP LEFT AND RUN  
| |  
LTURN WALK RUN --- LTURN JUMP RUN

Équivariance par permutation ?

Merci de votre attention

## Bibliographie (partielle)

- ▶ A. Cohan, F. Dernoncourt, D. S. Kim, T. Bui, S. Kim, W. Chang, N. Goharian. A discourse-aware attention model for abstractive summarization of long documents. NAACL 2018
- ▶ N. Craswell, B. Mitra, E. Yilmaz, D. Campos, E. M. Voorhees. Overview of the trec 2019 deep learning track. arXiv 2020
- ▶ M. Ding, C. Zhou, H. Yang, J. Tang. CogLTX : Applying BERT to Long Texts. NeurIPS 2020
- ▶ Y. Fang, S. Sun, Z. Gan, R. Pillai, S. Wang, J. Liu. Hierarchical graph network for multi-hop question answering. CoRR, abs/1911.03631, 2019
- ▶ Q. Grail, J. Perez, E. Gaussier. Globalizing BERT-based Transformer Architectures for Long Document Summarization. EACL 2021
- ▶ M. Joshi, O. Levy, L. Zettlemoyer, D. S. Weld. BERT for coreference resolution : Baselines and analysis. EMNLP-IJCNLP 2019
- ▶ C. Kedzie, K. McKeown, H. Daume III. Content selection in deep learning models of summarization. EMNLP 2018
- ▶ N. Kim, T. Linzen. COGS : A Compositional Generalization Challenge Based on Semantic Interpretation. EMNLP 2020
- ▶ B. M. Lake, M. Baroni. Generalization without Systematicity : On the Compositional Skills of Sequence-to-Sequence Recurrent Networks. ICML 2018

## Bibliographie (partielle)

- ▶ C. Li, A. Yates, S. MacAvaney, B. He, Y. Sun. PARADE : Passage representation aggregation for document reranking. arXiv 2020
- ▶ M. Li, E. Gaussier. KeyBLD : Selecting Key Blocks with Local Pre-ranking for Long Document Information Retrieval. SIGIR 2021
- ▶ M. Tu, K. Huang, G. Wang, J. Huang, X. He, B. Zhou. Select, answer and explain : Interpretable multi-hop reading comprehension over multiple documents. CoRR, abs/1911.00484, 2019
- ▶ Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, E. Hovy. 2016. Hierarchical attention networks for document classification. NAACL 2016