Une Analyse du Modèle ColBERT

Thibault Formal[*,] — Benjamin Piwowarski[*,***] — Stéphane Clinchant [**]**

[*] *LIP6 - Sorbonne Université, UPMC Univ Paris 06, UMR 7606*
benjamin.piwowarski@lip6.fr
[**] *Naver Labs Europe, Meylan, France*
{thibault.formal,stephane.clinchant}@naverlabs.com
[***] *CNRS*

RÉSUMÉ. *Les modèles de RI basés sur les Transformers sont aujourd'hui état de l'art en Recherche d'Information ad-hoc, mais leur comportement reste encore incompris. Des travaux récents ont montré que BERT ne satisfait pas les axiomes classiques de la RI. Nous proposons d'étudier le processus d'appariement du modèle ColBERT par l'analyse de l'importance des termes et des mécanismes d'appariement exact et sémantique. Même si les axiomes classiques ne sont pas formellement vérifiés, notre analyse révèle que ColBERT: (i) inclut implicitement une notion d'importance des termes; (ii) s'appuie sur des correspondances exactes pour les termes importants.*

ABSTRACT. *Transformer-based models are nowadays state-of-the-art in adhoc Information Retrieval, but their behavior is far from being understood. Recent work has claimed that BERT does not satisfy the classical IR axioms. However, we propose to dissect the matching process of ColBERT, through the analysis of term importance and exact/soft matching patterns. Even if the traditional axioms are not formally verified, our analysis reveals that ColBERT (i) is able to capture a notion of term importance; (ii) relies on exact matches for important terms.*

MOTS-CLÉS : *Transformers, BERT, Poids des termes.*

KEYWORDS: *Transformers, BERT, Term weights.*

Cet article a été accepté à ECIR 2021 en tant que papier court (Formal et al.*, 2021) (**best short paper award**). Nous présentons ci-dessous un résumé étendu.*

1. Introduction

Au cours des deux dernières années, le Traitement du Langage Naturel a été bouleversé par la publication de modèles de langue pré-entraînés basés sur l'attention propre (*self-attention*), e.g. BERT (Devlin *et al.*, 2018). Les modèles d'ordonnancement basés sur BERT sont actuellement état de l'art en RI ad-hoc, sur les *leaderboards* tels que MS MARCO (Nogueira et Cho, 2019) – dont ils occupent les premières places – aux jeux de données RI plus classiques comme Robust04 (MacAvaney *et al.*, 2019 ; Dai et Callan, 2019 ; Nogueira *et al.*, 2020). Il est de ce fait intéressant de mieux comprendre le fonctionnement de ces modèles. Certains travaux ont été menés dans ce sens (Rennings *et al.*, 2019 ; Câmara et Hauff, 2020), mais se sont concentrés sur le respect – ou non – d'axiomes de RI. Dans (Câmara et Hauff, 2020), il a ainsi été montré qu'un modèle de re-ordonnancement basé sur BERT ne respecte pas entièrement certains axiomes jugés importants pour les modèles RI standards, comme l'axiome indiquant que les *mots apparaissant dans un plus grand nombre de documents sont moins importants* (effet IDF). D'autres études se sont écartées de cette approche axiomatique, comme (MacAvaney *et al.*, 2020) qui s'intéresse à certaines propriétés du langage comme l'ordre des mots ou la fluidité du texte. Mais ces deux types de travaux ne permettent pas vraiment de comprendre *comment* fonctionnent ces modèles.

Il existe une grande variété de modèles d'ordonnancement basés sur BERT (Lin *et al.*, 2020). Les modèles BERT *standards* (basés sur un encodage joint de la question et du document) sont difficiles à analyser car ils nécessitent une analyse approfondie des mécanismes d'attention, qui se révèle complexe (Brunner *et al.*, 2020). De ce fait, nous avons plutôt choisi de nous concentrer sur les modèles d'interaction contextuelle, pour lesquels la question et le document sont encodés indépendamment (MacAvaney *et al.*, 2019 ; Hofstätter *et al.*, 2020 ; Khattab et Zaharia, 2020). Parmi ces modèles, ColBERT (Khattab et Zaharia, 2020) se distingue pour deux raisons principales : (i) il offre le meilleur compromis efficacité/efficience ; (ii) sa fonction de score est une simple somme sur les sous-tokens de la question[1], ce qui le rend comparable aux modèles RI standards comme BM25, et facilite l'analyse, puisque la contribution de chaque terme est *explicite*.

Dans cet article, nous nous concentrons donc sur ColBERT et examinons deux questions de recherche. Plus précisément, nous étudions dans un premier temps le lien entre l'importance des termes telle que calculée par les modèles RI standards, et celle calculée par ColBERT (**RQ1**). Ensuite, nous examinons comment ColBERT traite les correspondances exactes (même mot) et sémantiques (mot sémantiquement lié) (**RQ2**). En particulier, nous montrons (via une une analyse spectrale) que le mo-

1. Comme tous les modèles Transformers, la segmentation utilisée par ColBERT est une segmentation apprise de manière non supervisée, et qui peut séparer un mot en plusieurs sous-mots.

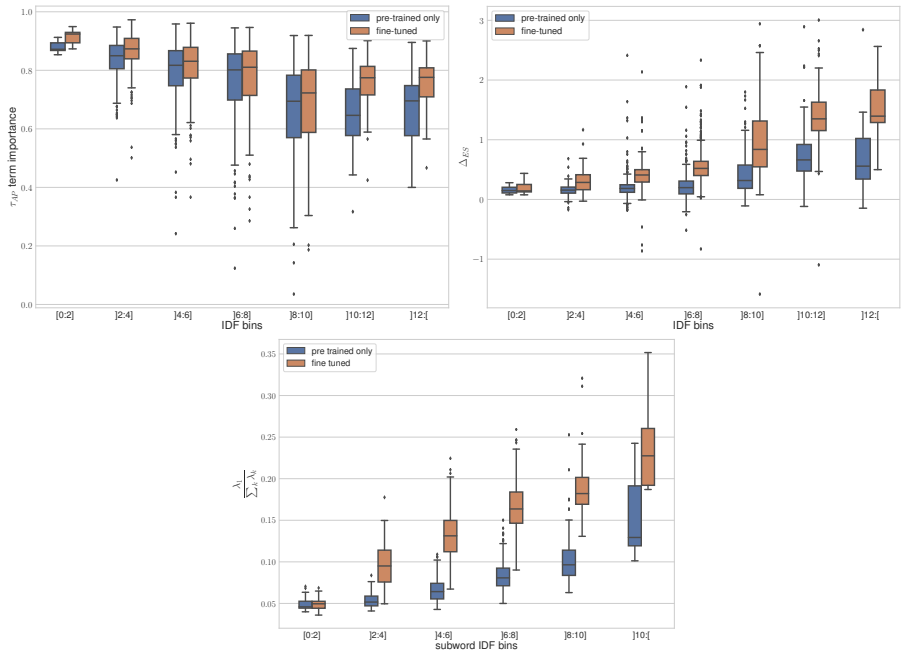

Figure 1. *(a) Importance des termes pour ColBERT (τ_{AP}) en fonction de l'IDF ; (b) Δ_{ES} en fonction de l'IDF ; (c) $\frac{\lambda_1}{\sum_i \lambda_i}$ en fonction de l'IDF (au niveau du sous-token).*

dèle favorise les appariements exacts pour les termes importants (IDF élevé), pour lesquels les représentations contextualisées varient peu. Pour notre analyse, nous considérons le jeu de données MS MARCO, et nous nous plaçons dans un cadre de *réordonnancement*, où, pour chaque question, le modèle doit reclasser un ensemble de documents sélectionnés par un premier modèle (e.g. BM25).

2. Importance des termes pour ColBERT

Notre première question de recherche (**RQ1**) porte sur la comparaison de l'importance des termes dans les modèles RI standards (par exemple BM25) avec l'importance des termes telle que déterminée par ColBERT. Pour ce dernier, il est difficile de mesurer l'importance d'un terme, car elle dépend à la fois du contexte du document et de la question. Nous avons donc recours à une approche indirecte, qui mesure la corrélation entre le classement des documents initialement induit par ColBERT, et le classement obtenu lorsque nous retirons toutes les *contributions* des sous-tokens qui composent le terme correspondant. Pour comparer ces listes ordonnées de documents, nous utilisons la corrélation AP τ_{AP} (Yilmaz *et al.*, 2008). Une valeur proche de 1 signifie que les deux classements sont similaires, impliquant une faible contribution du

terme dans le processus d'ordonnancement – et donc une faible *importance*. Il existe une corrélation linéaire négative entre cette mesure et l'IDF du terme correspondant (Figure 1(a), coefficient de corrélation de Pearson r=-0.4) : ColBERT capture ainsi implicitement une notion d'importance des termes. Il est à noter que les termes ayant un IDF plus élevé ont tendance à être plus longs, et donc à être plus souvent divisés en plusieurs sous-tokens, augmentant d'autant plus leur importance.

3. Analyse des correspondances exactes et sémantiques

Après avoir examiné l'importance des termes, nous nous sommes intéressés à la façon dont les correspondances exactes sont traitées par ColBERT (**RQ2**). Pour cela, nous avons mesuré la différence entre l'*importance* moyenne d'un terme de la question lors d'un appariement exact (ColBERT sélectionne le mot de la question dans le document) ou sémantique (ColBERT sélectionne un mot qui n'est pas celui de la question). Nous avons observé que cette mesure (Δ_{ES}) est une nouvelle fois corrélée avec l'IDF (Figure 1(b), r=0.667). Il est intéressant de noter que cet effet est déjà observable lorsque le modèle n'a pas été entraîné pour la tâche de RI, mais que l'ajustement (ou *fine-tuning*) du modèle sur des données de pertinence a un impact important pour les mots dont l'IDF est supérieur à 8, pour lesquels ColBERT apprend à mettre l'accent sur les correspondances exactes.

Pour expliquer ce comportement, notre hypothèse est que les correspondances exactes sont dues à des représentations contextuelles qui varient peu : dans ce cas, la similarité entre le terme de la question et le terme du document serait plus proche de 1, et ColBERT aura tendance à *sélectionner* systématiquement ce terme. Au contraire, les termes qui véhiculent moins d'"information" sont plus fortement influencés par leur contexte ; ainsi, leur représentation a tendance à plus varier. Pour vérifier cette hypothèse, nous avons procédé à une analyse spectrale (décomposition en valeurs singulières) des représentations contextuelles d'un même terme dans l'ensemble des documents où il apparaît. Si la représentation ne varie pas – ou presque pas – une valeur singulière devrait se démarquer : nous avons donc calculé le ratio entre la valeur singulière de plus haute magnitude et la somme de toutes les valeurs singulières. Ce ratio augmente avec l'IDF (Figure 1(c), r=0.77), ce qui valide notre hypothèse. De plus, cet effet est beaucoup plus fort après le *fine-tuning* du modèle, indiquant que l'apprentissage sur une tâche de RI favorise les appariements exacts dans ColBERT.

4. Conclusion

Dans (Formal *et al.*, 2021), nous avons mis en évidence (i) que même si l'effet IDF de la théorie axiomatique n'est pas formellement vérifié, le modèle ColBERT capture une notion d'importance du terme ; (ii) que la correspondance exacte entre termes reste une composante critique du modèle, en particulier pour les termes importants ; (iii) que cet effet est lié aux propriétés des mots importants qui varient peu dans l'espace des représentations.

5. Bibliographie

Brunner G., Liu Y., Pascual D., Richter O., Ciaramita M., Wattenhofer R., « On Identifiability in Transformers », *ICLR*, 2020.

Câmara A., Hauff C., « Diagnosing BERT with Retrieval Heuristics », *in* J. M. Jose, E. Yilmaz, J. Magalhães, P. Castells, N. Ferro, M. J. Silva, F. Martins (eds), *Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part I*, vol. 12035 of *Lecture Notes in Computer Science*, Springer, p. 605-618, 2020.

Dai Z., Callan J., « Deeper Text Understanding for IR with Contextual Neural Language Modeling », *CoRR*, 2019.

Devlin J., Chang M., Lee K., Toutanova K., « BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding », *CoRR*, 2018.

Formal T., Piwowarski B., Clinchant S., « A White Box Analysis of ColBERT », *ECIR*, 2021.

Hofstätter S., Zlabinger M., Hanbury A., « Interpretable & Time-Budget-Constrained Contextualization for Re-Ranking », *in* G. D. Giacomo, A. Catalá, B. Dilkina, M. Milano, S. Barro, A. Bugarín, J. Lang (eds), *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)*, vol. 325 of *Frontiers in Artificial Intelligence and Applications*, IOS Press, p. 513-520, 2020.

Khattab O., Zaharia M., « ColBERT : Efficient and Effective Passage Search via Contextualized Late Interaction over BERT », *ACM SIGIR*, SIGIR '20, Association for Computing Machinery, New York, NY, USA, p. 39–48, 2020.

Lin J., Nogueira R., Yates A., « Pretrained Transformers for Text Ranking : BERT and Beyond », *arXiv :2010.06467 [cs]*, October, 2020.

MacAvaney S., Feldman S., Goharian N., Downey D., Cohan A., « ABNIRML : Analyzing the Behavior of Neural IR Models », *arXiv :2011.00696 [cs]*, 2020.

MacAvaney S., Yates A., Cohan A., Goharian N., « CEDR : Contextualized Embeddings for Document Ranking », *SIGIR*, 2019.

Nogueira R., Cho K., « Passage Re-ranking with BERT », *arXiv :1901.04085 [cs]*, 2019.

Nogueira R., Jiang Z., Lin J., « Document Ranking with a Pretrained Sequence-to-Sequence Model », *arXiv :2003.06713 [cs]*, 2020.

Rennings D., Moraes F., Hauff C., « An Axiomatic Approach to Diagnosing Neural IR Models », *in* L. Azzopardi, B. Stein, N. Fuhr, P. Mayr, C. Hauff, D. Hiemstra (eds), *Advances in Information Retrieval*, Lecture Notes in Computer Science, Springer International Publishing, Cham, p. 489-503, 2019.

Yilmaz E., Aslam J. A., Robertson S., « A New Rank Correlation Coefficient for Information Retrieval », *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, Association for Computing Machinery, New York, NY, USA, p. 587–594, 2008.