A White Box Analysis of ColBERT

Thibault Formal[1,2](✉), Benjamin Piwowarski[1], and Stéphane Clinchant[2]

[1] LIP6, Sorbonne Université, 75005 Paris, France
benjamin.piwowarski@lip6.fr
[2] Naver Labs Europe, Meylan, France
{thibault.formal,stephane.clinchant}@naverlabs.com

Abstract. Transformer-based models are nowadays state-of-the-art in adhoc Information Retrieval, but their behavior are 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.

Keywords: Information retrieval · Term matching · Transformer · BERT

1 Introduction

Over the last two years, Natural Language Processing has been shaken by the release of large pre-trained language models based on self-attention, like BERT [4]. Ranking models based on BERT are currently state-of-the-art in adhoc IR, ranking first on leaderboards[1] of the MSMARCO passage and document (re-)ranking tasks by a large margin [11], as well as on more standard IR datasets such as Robust04 [3,10,12]. It is thus interesting to understand better what is happening inside those models, and what phenomena are captured. Some works have been conducted in this direction [2,13], but focused on whether IR axioms are respected – or not – by neural and transformer-based models. In [2], BERT has been shown to not fully respect axioms that have proved to be important for standard IR models, such as the axiom stating that words occurring in more documents are less important (IDF effect). [9] extended the diagnosis to properties like word order or fluency. Instead of investigating whether these models behave like standard ones, we make a step towards understanding *how* they manage to improve over traditional models through their specific matching process.

There exists a wide variety of BERT-based ranking models, as summarized in the recent overview [8]. Canonical BERT models are difficult to analyse because they require a thorough analysis of attention mechanisms, which is a complex

[1] https://microsoft.github.io/msmarco/.

D. Hiemstra et al. (Eds.): ECIR 2021, LNCS 12657, pp. 257–263, 2021.
https://doi.org/10.1007/978-3-030-72240-1_23

task [1]. We rather choose to focus on contextual interaction models [6,7,10], where query and document are encoded *independently*. Among such models, ColBERT [7] exhibits the best trade-off between effectiveness and efficiency, with performance on par with standard BERT, suggesting that the power of these models comes from learning rich contextual representations, rather than modeling complex matching patterns. Moreover, the structure of ColBERT (sum over query terms of some similarity scores) is similar to standard IR models like BM25, and makes the analysis easier, as the contribution for each term is explicit.

In this paper, we hence focus on ColBERT, and look at two research questions. In Sect. 3, we investigate the link between term importance as computed by standard IR models, and the one computed by (Col)BERT. In Sect. 4, we look at how (Col)BERT is dealing with exact and soft matches as this is known to be critical for IR systems.

2 Experimental Setting

Dataset. For our analysis, we use the passage retrieval tasks from TREC-DL 2019 and 2020 [15] (400 queries in total). We consider a re-ranking setting, where for a given query q, the model needs to re-rank a set of documents $\mathcal{S}_q$ selected by a first stage ranker. Following the MSMARCO setting, we consider candidates from BM25, and $|\mathcal{S}_q| \leq 1000$. In order to study the model properties, we are interested in *how it attributes scores to each query token, for documents in* $\mathcal{S}_q$.

ColBERT. We now introduce the variant of ColBERT [7] we used to simplify the analysis – we checked each time that the drop in performance was minor. In particular, we did not include query/document specific tokens, since they could bias the term representations. Second, while query augmentation has been shown to be beneficial in [5,7], we omit this component to avoid the analysis of the induced implicit query expansion mechanism. We however keep the compression layer, that projects token representations from the BERT space ($d = 768$) to the ColBERT space ($d = 128$). By fine-tuning our model in a similar fashion to [7], we obtain a MRR@10 of 0.343 on MSMARCO dev set (versus 0.349). This shows that the above simplifications are negligible performance-wise, and would not invalidate our analysis. In order to understand what is learned during training, we also consider a non fine-tuned version of the model (without compression layer), that relies on the output of a pre-trained BERT model.

The formal definition of ColBERT, given the BERT embeddings $E_q = (E_{q_i})_i$ for the query q (after WordPiece tokenization) and $E_d = (E_{d_j})_j$ for the document d, is given by the following relevance score:

$$s(q,d) = \sum_{i \in q} \max_{j \in d} \cos(E_{q_i}, E_{d_j}) = \sum_{i \in q} \max_{j \in d} C_{ij} = \sum_{i \in q} C^{\star}_{id} \tag{1}$$

In the following, we say that *a query token i matches the document token j^** if $C_{ij^*} = C^{\star}_{id}$. We denote this token j^* by $d^{\star}_i$.

3 ColBERT Term Importance

Our first research question focuses on comparing term importance in standard IR models (e.g. BM25) with term importance as determined by ColBERT. With respect to the former, given that documents are small passages, term frequency is close to 1 for most terms ($\text{avg}(tf) \approx 1.1$). Moreover, passage length does not vary much, and is caped at 512 tokens. Hence, we can reasonably assume that a term BM25 score roughly corresponds to its IDF – this might not be true for terms with low IDF, but it is a good enough approximation for other terms.

For ColBERT, it is difficult to measure the importance of a term, as it depends on both document and query contexts. We hence resort to an indirect mean, by measuring the correlation between the original ColBERT ranking and the ranking obtained when we remove from the sum in Eq. (1) all the *contributions* of subwords that compose the corresponding term. Another option would be to directly mask the input term, but we would loose the query structure. Finally, to compare rankings, we use AP-correlation[2] τ_{AP} [16], which is akin to Kendall rank correlation, but gives more importance to the top of the ranking. Values close to 1 indicate a strong correlation, meaning that the two rankings are similar, implying a low contribution of the term in the ranking process. Note that such measure of importance is query dependent: when the term appears in several queries, we consider the average as a final measure of importance.

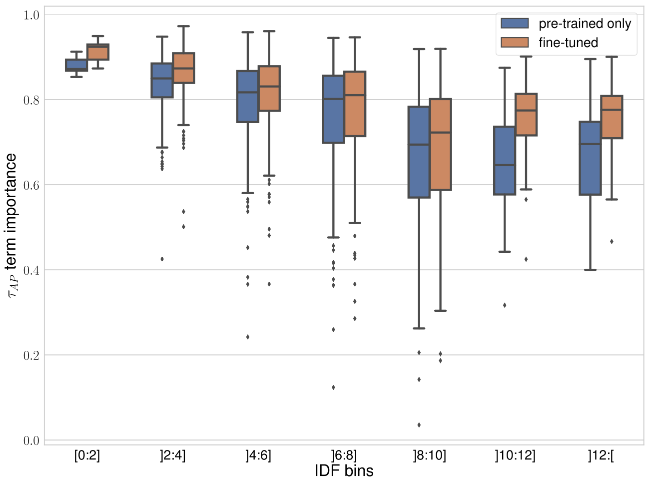

Fig. 1. ColBERT term importance (as computed using τ_{AP}) with respect to IDF.

In Fig. 1, we show how IDF and τ_{AP} are connected. There is a linear negative correlation between both metrics (Pearson correlation coefficient $r = -0.4$), showing that (Col)BERT implicitly captures IDF. Note that words with higher

[2] using the Python implementation provided by [14].

IDF tend to be longer, and hence to be split into multiple subwords more often – increasing the importance of such terms.

We also observe that the link between IDF and term importance is not so direct for high IDF values (>8). We believe that there are three reasons explaining this behavior: (i) ColBERT has correctly learned that this term was not *so* important; (ii) as most of the documents contain the term, the effect on τ_{AP} might not be high; (iii) another query term (with no semantics) is bearing the same semantics as the target one. The first hypothesis is probably true since ColBERT improves over BM25. As for the second one, this is a more general observation regarding the re-ranking setting, where IR axioms might not fully apply. Finally, to investigate (iii), we looked, for each query token, at the frequency of exact matching (i.e. the max similarity is obtained with the same token in a document) and at the frequency with which it matches in documents *other* query terms. We observed that stopwords (*the*, *of*, etc.) did indeed match terms in the documents that were other query terms. For instance, in the query (and associated τ_{AP}) "*the (0.94) symptoms (0.87) of (0.93) shingles (0.88)*", the word "of" actually mostly matches with "shingles" in documents from $\mathcal{S}_q$.

4 Analysis of Exact and Soft Matches

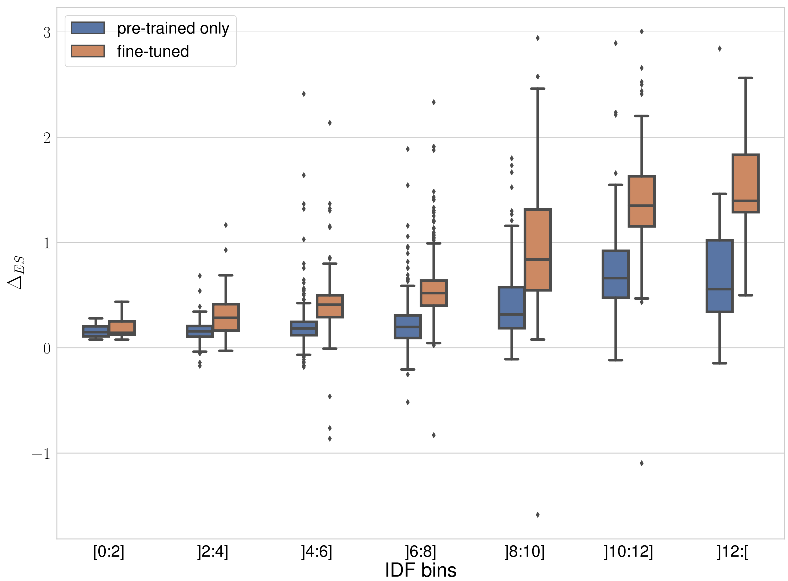

Fig. 2. Δ_{ES} with respect to IDF: we observe a moderate correlation (0.667), showing that the less frequent a term is, the more it is likely to be matched exactly.

After having looked at term importance, we now turn our attention into the issue of exact matches, i.e. how exact string matching is processed by ColBERT.

We need to define a measure indicating when ColBERT asserts whether a term should favor an exact match or not (i.e., soft match). To do so, we compute, for each query term i, the difference between the average ColBERT scores when i matches the same term within a document (i.e., when $d_i^\star \rightarrow t$) or not (i.e., when $d_i^\star \not\rightarrow t$). We then average at the query level, to obtain one measure per term (for terms appearing in several queries). This measure is formally defined as:

$$\Delta_{ES}(t) = \underset{i,q/i \rightarrow t}{\mathrm{mean}} \left(\underset{d \in \mathcal{S}_q/d_i^\star \rightarrow t}{\mathrm{mean}} \{C_{id}^\star\} - \underset{d \in \mathcal{S}_q/d_i^\star \not\rightarrow t}{\mathrm{mean}} \{C_{id}^\star\} \right) \tag{2}$$

where $j \rightarrow t$ means that the j^{th} token corresponds to token t.

For a term w composed of several WordPiece components $t_1, \ldots, t_n$, we use $\sum_{t \in w} \Delta_{ES}(t)$, which corresponds to the way ColBERT operates (summing over subwords). Then, for each query term w, we plot $\Delta_{ES}(w)$ with respect to $IDF(w)$ (Fig. 2). Higher Δ tends to indicate that a match value is higher if the terms appears in the document (exact match), as the model learns to widen the gap (in average) between exact and soft scores. We can observe a moderate positive correlation between terms focusing more on exact matching –larger Δ_{ES}– and IDF ($r = 0.667$). Interestingly, this effect is already observable for BERT, but fine-tuning has an important impact for words with an IDF above 8: ColBERT thus learns to emphasize on exact matches for such words. For instance, in the query (and associated Δ_{ES}) "*causes (0.35) of (0.11) left (0.64) ventricular (1.14) hypertrophy (1.62)*", the model mostly relies on exact match for the last two terms.

To explain this behavior, our hypothesis is that exact matches correspond to contextual embeddings that do not vary much: hence, the cosine similarity between the query term and the document term would be closer to 1, and ColBERT will tend to select this term. On the contrary, terms that carry less "information" are more heavily influenced by their context (they act as some sort of reservoirs to encode concepts of the sequence), and thus their embeddings vary a lot. To check this hypothesis, we conducted a spectral analysis of contextual term embeddings. More specifically, we use an SVD decomposition of the matrix composed of all the contextual representations for a given term t, on the test documents, and look at the relative magnitude of the singular values $\lambda_1 \geq \ldots \geq \lambda_d$ where d is the dimension of the embedding space. If the magnitude of λ_1 is much larger than the others, it means that all the contextual representations point to the same direction in the embedding space. In Fig. 3, we report the ratio of the first eigenvalue λ_1 with respect to $\sum_k \lambda_k$ for terms that appear in the test queries. It confirms the above hypothesis, as the ratio increases with the subword IDF (correlation $r = 0.77$). Moreover, this effect is much stronger when fine-tuning, indicating that training on relevance indeed promotes exact matches in ColBERT. By looking at the distribution of singular values (not shown here), we can confirm this trend. In particular, words with a low IDF tend to point in different directions, showing that what they capture is more about their context. For instance, in the query "*when did family feud come out ?*" (a TV show), the term "come", for all the documents in S_q, matches 97%

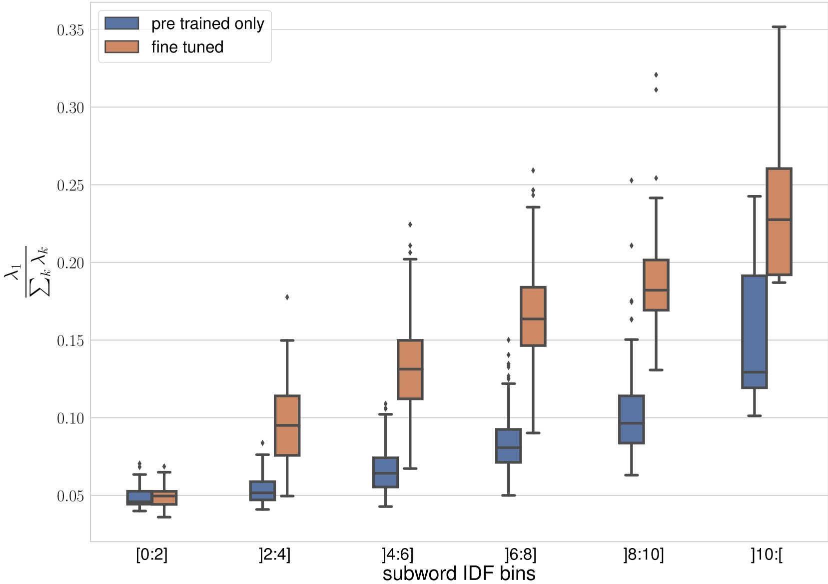

Fig. 3. Ratio of the first eigenvalue to the sum of the eigenvalues with respect to IDF (subword level). The less frequent the term is, the higher the ratio is, *showing that the contextualized embeddings for a rare term are concentrated in the same direction.*

of the time to terms that are not in the query, but are synonyms (in a broad sense) e.g. {*july, happen, item, landing, released, name, en, going, it, rodgers*}.

5 Conclusion

While the axiomatic approach is appropriate to analyze traditional IR models, its application to BERT-based models remains limited and somehow inadequate. To the best of our knowledge, our study is one of the first to shed light on matching behavior of BERT, through the analysis of a simpler counterpart, ColBERT. We showed that (i) even if the IDF effect from the axiomatic theory is not enforced, (Col)BERT does have a notion of term importance; (ii) exact matching remains an important component of the model, especially for important terms; (iii) our analysis gave some hints on the properties of frequent words which tend to capture the contexts in which they appear.

Although this work is a first step towards understanding matching properties of BERT in IR, we believe there is much more to uncover by either analyzing a wider range of models, or by extending our analysis of ColBERT to first stage ranking, where retrieval axioms might be more critical.

References

1. Brunner, G., Liu, Y., Pascual, D., Richter, O., Ciaramita, M., Wattenhofer, R.: On identifiability in transformers (2020)
2. Câmara, A., Hauff, C.: Diagnosing BERT with retrieval heuristics. In: Jose, J.M., et al. (eds.) ECIR 2020. LNCS, vol. 12035, pp. 605–618. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-45439-5_40
3. Dai, Z., Callan, J.: Deeper text understanding for IR with contextual neural language modeling. CoRR abs/1905.09217 (2019). http://arxiv.org/abs/1905.09217
4. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR abs/1810.04805 (2018). http://arxiv.org/abs/1810.04805
5. Hofstätter, S., Althammer, S., Schröder, M., Sertkan, M., Hanbury, A.: Improving efficient neural ranking models with cross-architecture knowledge distillation (2020)
6. Hofstätter, S., Zlabinger, M., Hanbury, A.: Interpretable & time-budget-constrained contextualization for re-ranking (2020)
7. Khattab, O., Zaharia, M.: 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 2020, pp. 39–48. Association for Computing Machinery, New York (2020). https://doi.org/10.1145/3397271.3401075
8. Lin, J., Nogueira, R., Yates, A.: Pretrained transformers for text ranking: BERT and beyond. arXiv:2010.06467 (October 2020)
9. MacAvaney, S., Feldman, S., Goharian, N., Downey, D., Cohan, A.: ABNIRML: analyzing the behavior of neural IR models (2020)
10. MacAvaney, S., Yates, A., Cohan, A., Goharian, N.: CEDR: contextualized embeddings for document ranking. In: SIGIR (2019)
11. Nogueira, R., Cho, K.: Passage re-ranking with BERT (2019)
12. Nogueira, R., Jiang, Z., Lin, J.: Document ranking with a pretrained sequence-to-sequence model (2020)
13. Rennings, D., Moraes, F., Hauff, C.: An axiomatic approach to diagnosing neural IR models. In: Azzopardi, L., Stein, B., Fuhr, N., Mayr, P., Hauff, C., Hiemstra, D. (eds.) ECIR 2019. LNCS, vol. 11437, pp. 489–503. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-15712-8_32
14. Urbano, J., Marrero, M.: The treatment of ties in AP correlation. In: Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2017, pp. 321–324. Association for Computing Machinery, New York (2017). https://doi.org/10.1145/3121050.3121106
15. Voorhees, E.M., Ellis, A. (eds.): Proceedings of the Twenty-Eighth Text REtrieval Conference, TREC 2019, Gaithersburg, Maryland, USA, November 13–15, 2019, vol. 1250. NIST Special Publication, National Institute of Standards and Technology (NIST) (2019). https://trec.nist.gov/pubs/trec28/trec2019.html
16. Yilmaz, E., Aslam, J.A., Robertson, S.: A new rank correlation coefficient for information retrieval. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, pp. 587–594. Association for Computing Machinery, New York (2008). https://doi.org/10.1145/1390334.1390435