Match Your Words! A Study of Lexical Matching in Neural Information Retrieval

Thibault Formal[1,2], Benjamin Piwowarski[2,3], Stéphane Clinchant[1]

[1] Naver Labs Europe
[2] Sorbonne Université, Institute for Intelligent Systems and Robotics
[3] CNRS

NAVER LABS Europe

Takeaways

Neural retrievers hold the promise to replace BM25 in modern search engines, but term matching still remains a critical component

We propose a **black-box approach** to measure a model ability to perform **lexical matching**, and answer the following questions:

(RQ1) To which extent neural retrievers capture lexical match (i.e. matching query terms) **when it's actually useful** (→ relevance)?

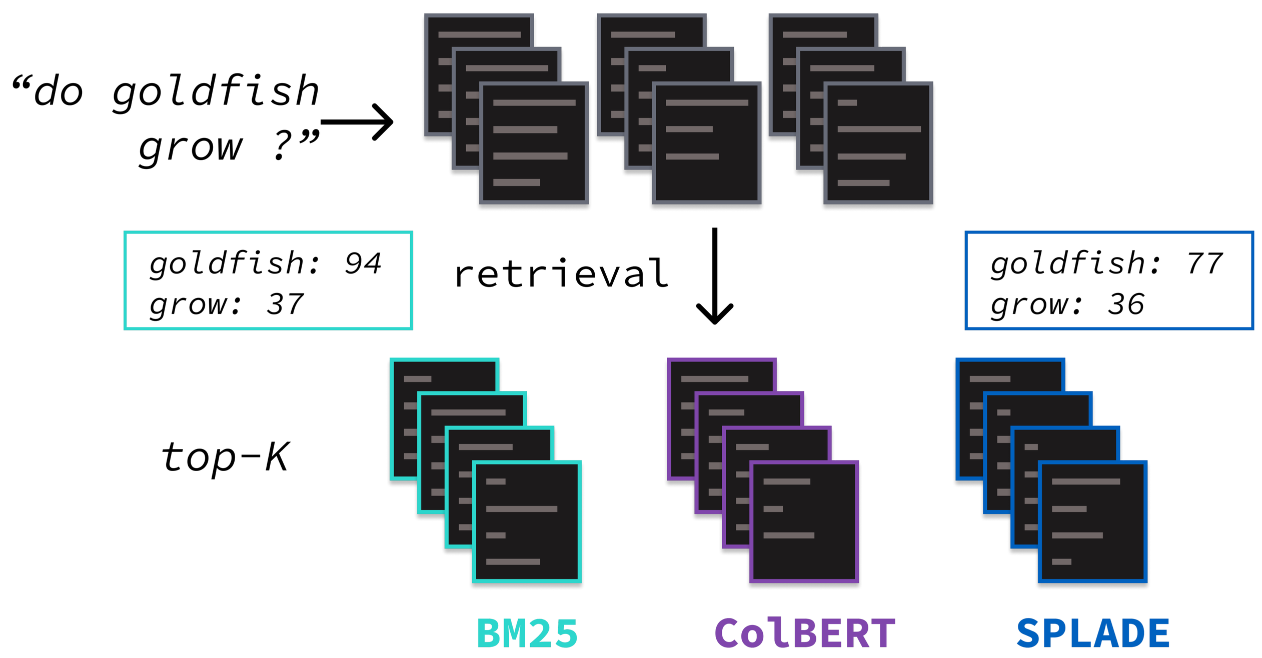
Do they generalize term matching to

- **(RQ2)** Terms not seen at training time?
- **(RQ3)** New collections?

Overall we show that neural IR models **fail to properly generalize term importance on out-of-domain collections or terms (almost) unseen at training time**

Method

High-level idea: "count" query terms in retrieved documents
Analysis rationale: the more a term is important for a query (w.r.t. relevant documents), the more a document containing it should be retrieved



Looking at frequency is not enough (e.g. stopwords): how to take into account **collection statistics + relevance?**

1. USER relevance (RSJ weight [1])

$$\mathrm{RSJ}_{t,U} = \log \frac{p(t|R)p(\neg t|\neg R)}{p(\neg t|R)p(t|\neg R)}$$

2. System relevance (derived from RSJ)

Hypothesis: top-K = documents considered to be relevant by the system

$$\mathrm{RSJ}_{t,S} = \log \frac{p(t|\text{top-}K)p(\neg t|\neg\text{top-}K)}{p(\neg t|\text{top-}K)p(t|\neg\text{top-}K)}$$

Contrast both values: look at $\Delta\mathrm{RSJ}=\Delta\mathrm{RSJ}_U-\Delta\mathrm{RSJ}_S$

- $\Delta > 0$: overestimates term importance
- $\Delta < 0$: underestimates term importance

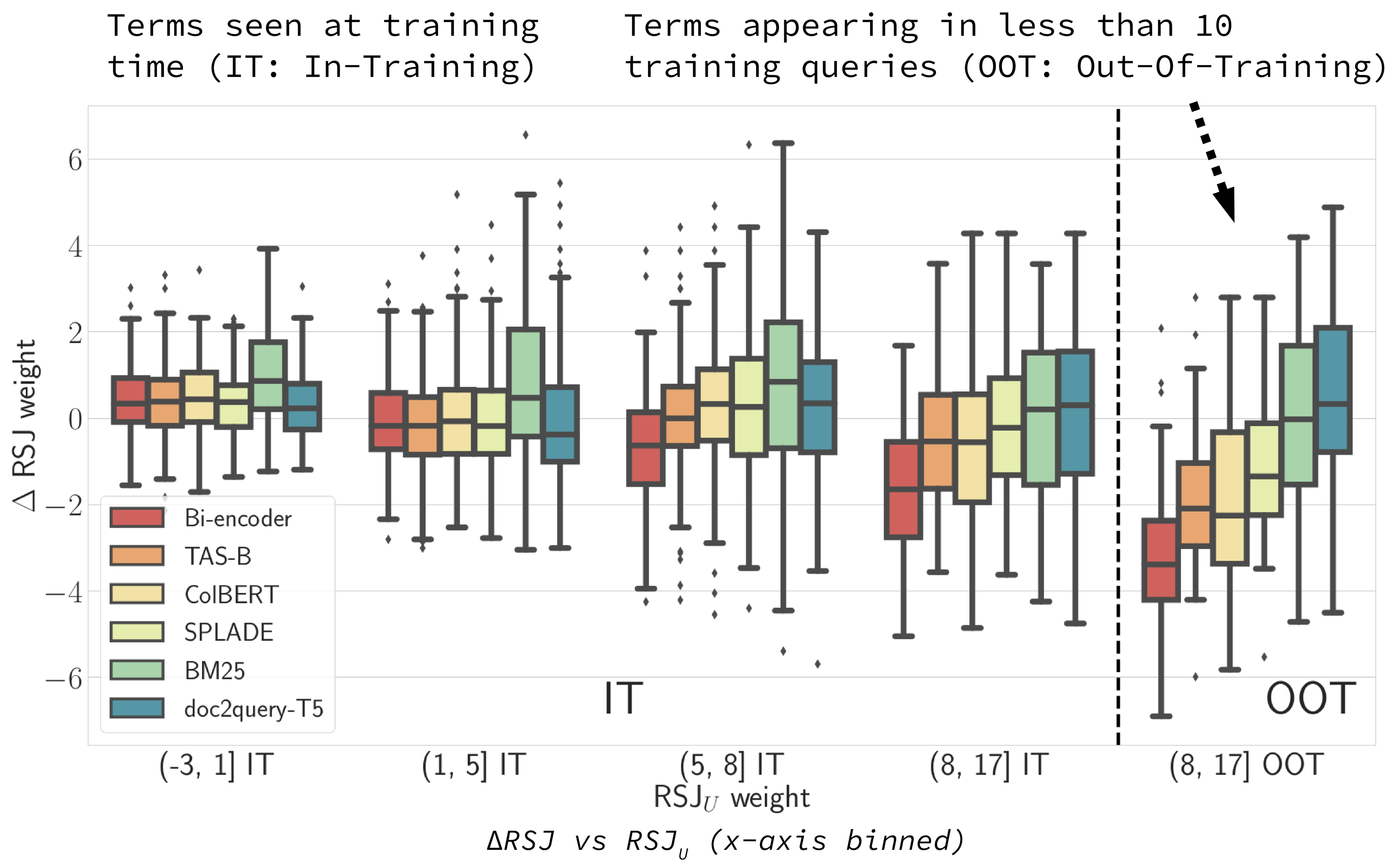
REFERENCES

[1] Relevance weighting of search terms, *S. E. Robertson, K. Sparck Jones*

[2] BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models, *Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, Iryna Gurevych*

In-domain / OOT terms

Evaluation on TREC 2019+2020 (97 queries)
Compare several dense and sparse neural models

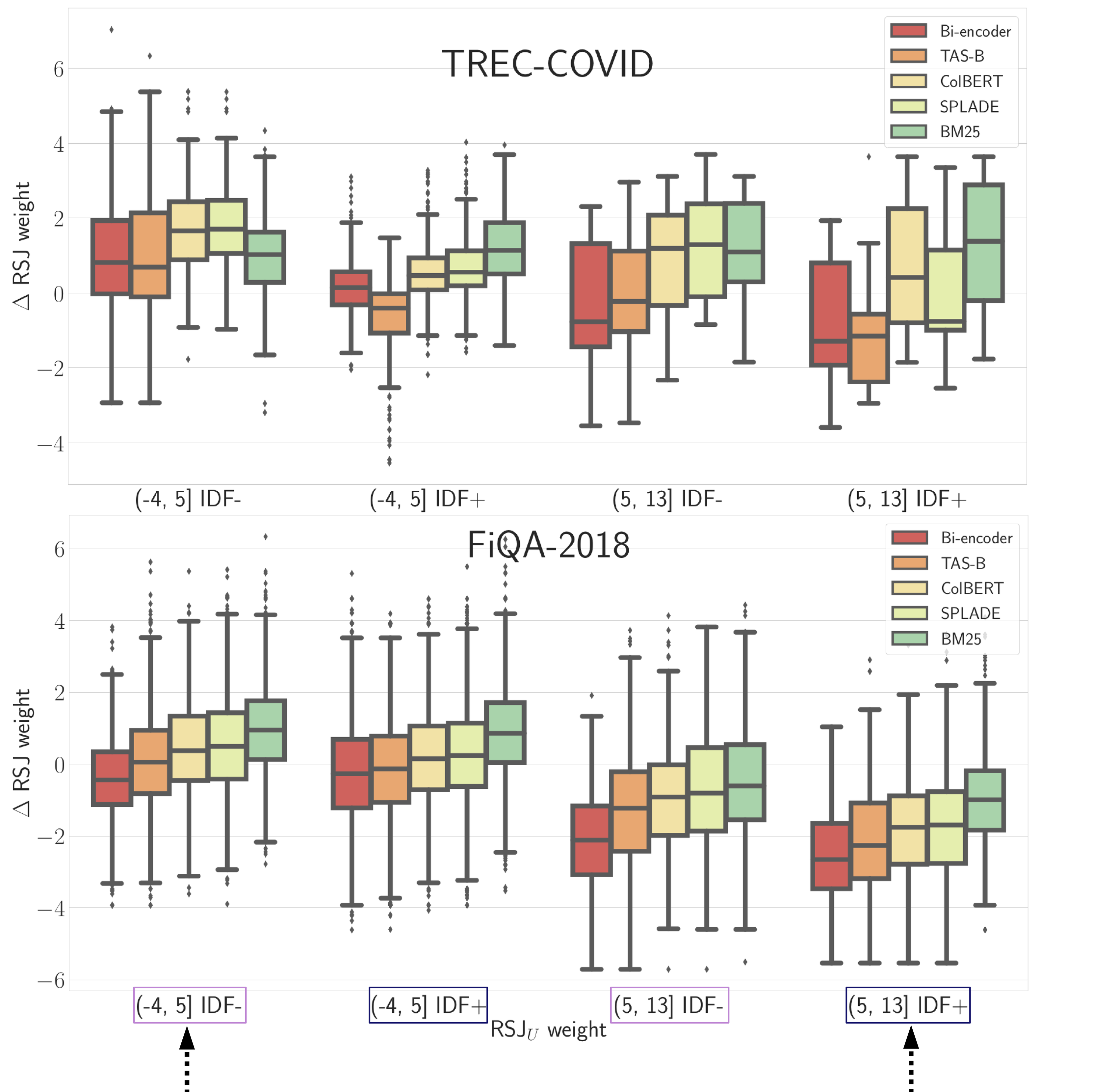


ΔRSJ vs RSJ_U (x-axis binned)

- "Order" between models (linked to lexical bias)
- For high RSJ, neural retrievers **underestimate** importance
- For unseen terms, it is worse

Out-of-domain

Evaluation on two out-of-domain datasets from the BEIR benchmark [2]: *TREC-COVID* and *FiQA-2018* (50 and 648 queries respectively)



IDF-: terms for which statistics are more or less unchanged

IDF+: terms which appear five times more in the new collection

- Overall, dense models underestimate while "sparse" ones tend to overestimate
- For terms with shifted statistics (IDF+), importance is underestimated
- Higher variance in Δ