From Distillation to Hard Negative Sampling: Making Sparse Neural IR Models More Effective

Thibault Formal
thibault.formal@naverlabs.com
Naver Labs Europe
Meylan, France
Sorbonne Université, ISIR
Paris, France

Carlos Lassance
carlos.lassance@naverlabs.com
Naver Labs Europe
Meylan, France

Benjamin Piwowarski
benjamin@piwowarski.fr
Sorbonne Université, ISIR, CNRS
Paris, France

Stéphane Clinchant
stephane.clinchant@naverlabs.com
Naver Labs Europe
Meylan, France

ABSTRACT

Neural retrievers based on dense representations combined with Approximate Nearest Neighbors search have recently received a lot of attention, owing their success to distillation and/or better sampling of examples for training – while still relying on the same backbone architecture. In the meantime, sparse representation learning fueled by traditional inverted indexing techniques has seen a growing interest, inheriting from desirable IR priors such as explicit lexical matching. While some architectural variants have been proposed, a lesser effort has been put in the training of such models. In this work, we build on SPLADE – a sparse expansion-based retriever – and show to which extent it is able to benefit from the same training improvements as dense models, by studying the effect of distillation, hard-negative mining as well as the Pre-trained Language Model initialization. We furthermore study the link between effectiveness and efficiency, on in-domain and zero-shot settings, leading to state-of-the-art results in both scenarios for sufficiently expressive models.

CCS CONCEPTS

• **Information systems → Retrieval models and ranking**.

KEYWORDS

neural networks, indexing, sparse representations, regularization

ACM Reference Format:
Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2022. From Distillation to Hard Negative Sampling: Making Sparse Neural IR Models More Effective. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3477495.3531857

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SIGIR '22, July 11–15, 2022, Madrid, Spain
© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3531857

1 INTRODUCTION

Traditional IR systems like BM25 have dominated search engines for decades [1], relying on lexical matching and inverted indices to perform efficient retrieval. Since the release of large Pre-trained Language Models (PLM) like BERT [4], Information Retrieval has witnessed a radical paradigm shift towards contextualized semantic matching, where neural retrievers are able to fight the long-standing vocabulary mismatch problem. In the first-stage ranking scenario, dense representations combined with Approximate Nearest Neighbors (ANN) search have become the standard approach, owing their success to improved training pipelines. While these models have demonstrated strong in-domain performance, their ability to generalize has recently been challenged on the recent zero-shot evaluation BEIR benchmark [26], where their average effectiveness is lower than BM25 on a set of various IR-related tasks.

In the meantime, there has been a growing interest in going back to the "lexical space", by learning sparse representations than can be coupled with inverted indexing techniques. These approaches, which generally learn term weighting and/or expansion, benefit from desirable IR priors such as explicit lexical matching and decades of works on optimizing the efficiency of inverted indices. These have also shown good generalization capabilities – w.r.t. either effectiveness [26] or IR behavior like exact match [5]. While they mostly differ in their architectural design, a lesser effort has been put in the training of such models, making it unclear how they would be able to take advantage of the same improvements as dense architectures. In this work, we build on the SPLADE model [6], and study the effect of adopting the latest advances for training dense retrievers. We provide an extensive experimental study – by training models in various scenarios – and illustrate the interplay between models capacity (as reflected by their sparsity) and performance. We show how improvements are additive, and how we are able to obtain state-of-the-art results for sufficiently expressive models.

2 RELATED WORKS

Replacing term-based approaches for candidate generation in search engine pipelines requires models and techniques that can cope with the high latency constraints of serving thousands of queries per