OSC at TREC 2020 - News track's Background Linking Task

Nathan Day Dan Worley Tim Allison
OpenSource Connections {nday@o19s.com}

Abstract

This notebook details a submission by OpenSource Connections (OSC) to the TREC 2020 News track's background linking task. OSC's strategies were built around Elasticsearch's More Like This (MLT) Query so they would be accessible to industry practitioners. OSC submitted four runs: MLT with default parameters, MLT with tuned parameters, MLT with a new named entity recognition (NER) field, and MLT re-scored by document embedding similarity. MLT with parameter tuning produced the best results of any OSC runs, but was slightly worse than the task median per topic of all submitted runs. OSC's NER and document embedding runs both produced better results than MLT with default parameters.

Introduction

News sources are critical to understanding current events. Often the events being covered are not isolated occurrences, they are parts of larger and more complex ones. The TREC 2020 News track is focused on providing news readers with better news context by linking to relevant related articles and Wikipedia resources. There are two distinct tasks in the News track, both operate identical input (a given news story) and evaluation metric (nDCG@5).

- Background Linking - Retrieve other news articles that provide important context or background information. The goal is to help a reader understand or learn more about the story or main issues in the current article using the best possible sources.
- Wikification - Identify short passages in the article that should be hyperlinked to either another article, or a Wikipedia article, in order to provide in-context access to information that would help contextualize or add background on the story being read.

The OSC team only participated in the Background Linking task.

Data

The data set for the News track is Washington Post articles from 2013-2017. There are 671,947 articles in the original dump, but after removing files per TREC News guidelines for relevant sections and wire service authorship, our final index had 615,727 documents.

Without the vector embedding required for OSC's *embed* run, the index was 3.5 Gb on disk. After including the document embeddings (768 floats), the index was 14.4 Gb on disk, a 314% increase compared to the original.

News articles were provided with their original sections intact, including title, kicker, and body paragraph delineations. News articles follow an inverted pyramid structure, where the importance of information decreases in subsequent paragraphs[1]. OSC's index attempted to take advantage of this intrinsic data, by using separate fields for the first three body paragraphs and a single field with all body paragraphs

collapsed together. Similar structure strategies have been employed in prior TREC News submissions[2] and we hypothesized placing more weight on earlier paragraphs would improve performance.

Approach

Elasticsearch, the most popular Lucene based search engine, was selected because of its specialized queries and widespread usage. Effort was focused on layering on improvements to Elasticsearch's out of the box capability by adjusting it's parameters, enriching documents with new fields for named entities and document embeddings. Each augmentation is represented as a run, and is detailed in the sub-sections below.

More Like This (base) Elasticsearch offers a specialized query, More Like This (MLT), for ranking similar documents from a given index. This query operates by selecting a representative set of terms, based on TF-IDF, for the original document. This set of terms is then used in a disjunctive (OR) query to recommend similar documents[3], making it a good candidate for the background linking task.

MLT allows parameter control of the total number of terms, the minimum/maximum term frequency for consideration, the minimum/maximum document frequency, the minimum/maximum word length and the fields considered.

Tuned MLT with Quaerite (tune) Because there are a plethora of parameters to adjust in MLT, the OSC team reached for the Quaerite toolkit[4]. Quaerite tries to improve search engine performance by searching for effective combinations of parameters and field weights, using techniques including genetic algorithms.

Ultimately the field weights identified by Quaerite were not included in the final run submitted to TREC, due to hypothesized over-fitting of the small (50 topics from TREC News 2018) training dataset. However, the parameter optimizations were used. The parameter changes from **base** MLT, involved increasing the maximum number of terms (from 25 to 50), the minimum word length (from 0 to 3) and minimum document frequency (from 5 to 10). The parameter for maximum word length was decreased (from unbounded to 20).

Named Entity Enrichment (ners) News stories are focused on describing events and events involve named entities. We hypothesized that isolating these entities into a new field could improve the performance of MLT. This idea has been utilized by past TREC News submissions[5].

To identify the entities OSC used the spaCY library[6] and its pre-trained en_core_web_lg model. Each document's title and body were submitted to the model. The identified entities were filtered to remove any entities tagged as CARDINAL, TIME, DATE, QUANTITY, or ORDINAL. This filter was hypothesized to reduce off-target matching, because quantitative measures could generate spurious matches in unrelated domains.

Document embeddings (embed) The vector representations BERT[7] produces are a big departure from the inverted indexes that dominate search engines, like Elasticsearch. Elasticsearch currently supports dense vector re-scoring via cosine similarity.

To generate embeddings for the index, we used Sentence-BERT (SBERT)[8]. SBERT is a BERT derivative specifically trained for semantic similarity search. We used the `distilbert-base-nli-stsb-mean-tokens` sentence transformer model included with the library. We generated a separate embedding for the first three body paragraphs before taking an average to represent the major topics of a given article.

Cosine similarity was used to calculate embedding similarity. This similarity was incorporated in the result ranking, via a re-score function performed after initial retrieval by MLT.

Results

OSC submitted four runs for the TREC News Background linking test, as outlined in the Methods sub-sections. Due to a technical error in the OSC evaluation script, the **ners** run was overwritten with the **tune** run's results. The **ners** run was evaluated unofficially after the competition results were scored, using the TREC evaluation scripts.

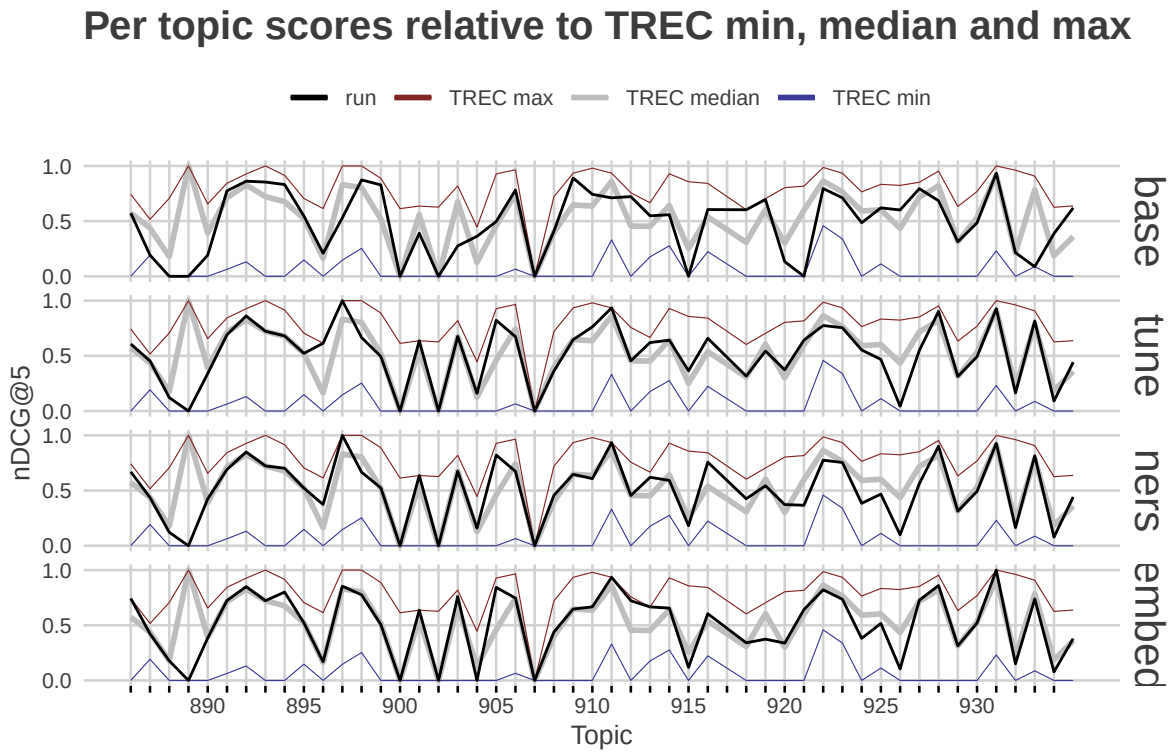
The News track's primary evaluation metric was nDCG@5, so we continue to use that metric for evaluating run performance. The only data available for comparison is the minimum, median, and maximum of all submitted runs on a per topic basis. There were 50 new topics in the 2020 evaluation set, numbers 886-934.

The table below reports the mean nDCG@5 score across all topics for each run and the TREC median. All of OSC's runs had lower mean nDCG@5 than the TREC median. None of OSC's runs were significantly different than the TREC median in a paired two-sample T-test.

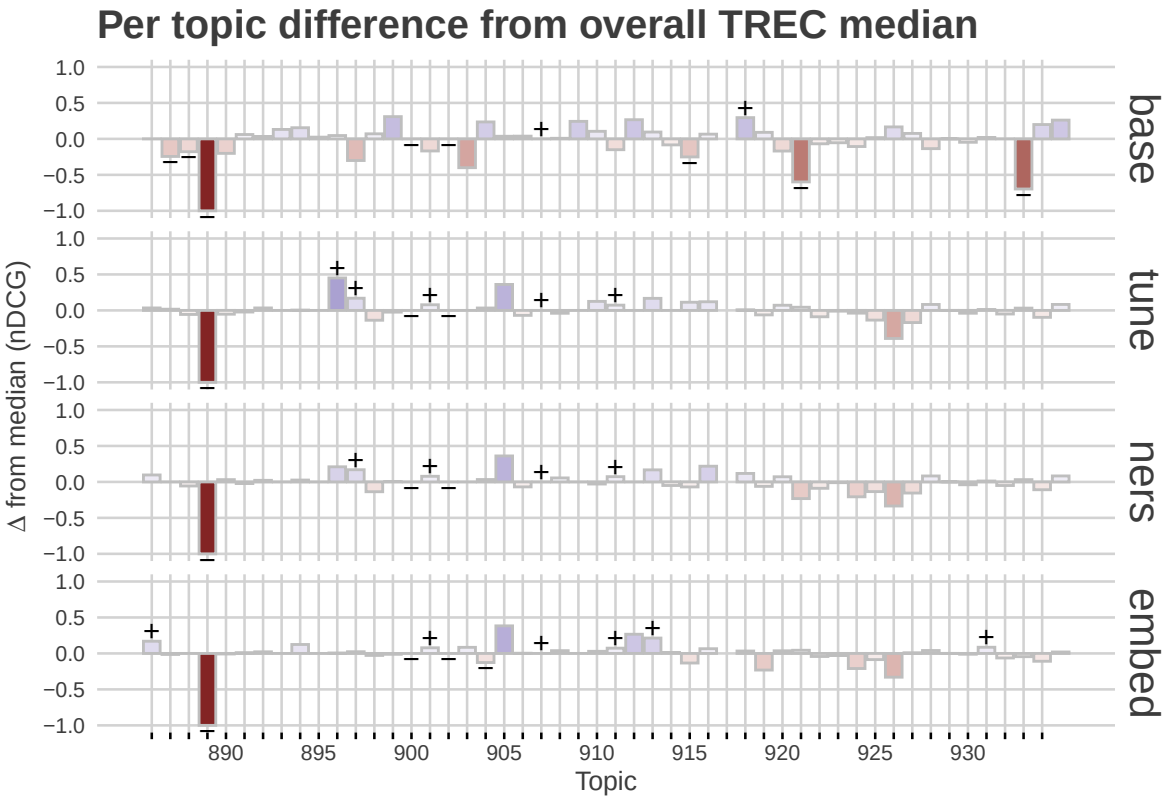
Run	Mean nDCG@5
base	0.488
tune	0.517
ners	0.506
embed	0.513
TREC median	0.525

The best performing OSC run was the simplest augmentation, tuning the parameters of MLT. The worst performing was the out of the box MLT.

In evaluating individual topic performance, topic 889 stands-out. All four OSC runs recorded a nDCG@5 score of 0.0, while the TREC median was 1.0. This visualization shows per topic nDCG@5 scores as a line and highlights each run's performance relative to the TREC minimum (min), median and maximum (max).



There were multiple topics where the scores of OSC runs tied with the TREC maximum score. This alternative visualization shows run performance per topic as the different from TREC median. Symbols denote when a run's per topic score tied the TREC maximum (+) or the TREC minimum (-).



The OSC **embed** run has the most ties (6) with the per topic maximums, while reporting a slightly lower mean that OSC's overall best **tune**.

Conclusion

OSC's best performing run was **tune** which is an adjusted parameter configuration of the MLT query. This was not surprising because the MLT query's default parameters are designed to be efficient as well as effective. In real world search applications query run time is important. The parameter settings of **tune** allow for a maximum of 50 terms instead of the 25 allowed in **base**. This potentially doubles the queries required for MLT, but for this accuracy focused competition we ignored this potential downside. If **tune** has scored the median (1.0) for topic 889, it's mean nDCG@5 score would be 0.537, slightly better than the mean of the TREC median per topic.

Neither OSC's named entity enrichment, **ners**, or document embedding similarity re-score **embed** was able to best the performance of **tune**, which was surprising. In testing approaches on the data from the the 2018 News track, OSC saw an increase in average performance in both **ners** and **embed**, with **embed** producing the best results on historical data.

Of interest for future competitions is using vector similarity for initial retrieval, not just re-ranking. Today this is not supported by Elasticsearch, or it's underpinning low level library Lucene, but there is active work for adding approximate nearest neighbor search. These developments represent a novel matching strategy for search engines and we are excited to explore their performance in future conferences.

Acknowledgments

The authors would like to thank Samantha Toet and Charlie Hull for their editorial feedback.

References

[1] News style. (2020, October 31). Retrieved November 09, 2020, from https://en.wikipedia.org/wiki/News_style

[2] Sumanta Kashyapi, Shubham Chatterjee, Jordan Ramsdell and Laura Dietz. TREMA-UNH at TREC 2018: Complex Answer Retrieval and News Track, 2018.

[3] More like this query: Elasticsearch Reference [7.9]. (2020). Retrieved November 09, 2020, from https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-mlt-query.html

[4] Allison, T. (2020, April). Quaerite (Version 1.0.0-Alpha2) [Computer software]. Retrieved June, 2020, from https://github.com/tballison/quaerite

[5] Agra Bimantara, Michelle Blau, Kevin Engelhardt, Johannes Gerwert,Tobias Gottschalk, Philipp Lukosz, Shenna Piri, Nima Saken Shaft and Klaus Berberich.htw saar @ TREC 2018 News Track, 2018

[6] Honnibal, M., & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Un-derstanding.arXiv preprint arXiv:1810.04805.

[8] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks.arXiv preprint arXiv:1908.10084, 2019.