Notebook for the Touché Lab on Argument Retrieval at CLEF 2021

Danik Hollatz, Philipp Reinhardt and Artem Chepurniak

Abstract

We report our efforts to employ retrieval models as part of an information retrieval pipeline, using argument retrieval as a benchmark. We investigated one approach to rank documents and four approaches to predict combined quality of premises in documents to use it later for re-ranking. In particular: ranking documents via tf-id methodoly and re-ranking them using GloVe [1], POS n-grams and BERT embeddings [2].

1. Introduction

Students of Martin Luther University as a part of lecture 'Web-searching and Information Retrieval" were given a task to write a program that can rank documents of a given corpus. Afterward that program had to be improved either via Query Expansion or, in our case, re-ranking using Machine Learning methods. The first part of our work was to create a baseline to rank documents of args.me [3] corpus. The Second part of our work was to train a regression model on a new corpus - Webis-ArgQuality-20 [4], which contains 1610 annotaed premises and corresponding Combined Quality that later has to be predicted and used for re-ranking of already retrieved documents. We also randomly split the dataset to 80% and 20% for train and test sets correspondingly.

2. Approaches

Before applying any approaches, we lowered all words in premises and deleted all stopwords listed in nltk [5]

2.1. TF-IDF

Tf-idf is an information retrieval technique that weighs a term's frequency (tf) and its inverse document frequency (idf). Each word or term that occurs in the text has its respective tf and idf score. The product of the tf and idf scores of a term is called the tf-idf weight of that term. To limit the bias towards larger documents, tf gets replaced by its logarithm. This methodology was picked by us as a baseline.

2.2. GloVe

GloVe is an unsupervised learning algorithm for obtaining vector representations of words. Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space. We used the-pretrained embeddings with 50 dimensions and created for each premise a normalized vector.

2.3. GloVe and POS bigrams

POS stands for Part of Speech. We have created embeddings of sentences containing POS bi- and unigramms. Those vectors possess some useful information. By expanding GloVe vectors with POS-tags, the regression models is able to give better predictions.

2.4. BERT Transformers

One of our approaches was also to create embeddings using third library 'transformes' and their pretrained BERT model. Embeddings created using this library were then fed to the regressor.

3. Results

The evaluation of our approaches to argument retrieval is as part of the Touché shared task.

3.1. Experimental Setup

We have tried different approaches to predict combined quality of a premise using an implemented regression model from XGBoost [6]. Following table show root mean squared error of each approach.

Model	RMSE
GloVe	1.48
GloVe and POS bigrams	1.31
BERT embeddings	1.27
POS bi- and unigrams	**1.19**

For evaluating the Touché shared task on argument retrieval the evaluation platform on TIRA uses trec-eval to judge entries to the competition. Every participating team have their own virtual machine where their retrieval models will be tested. Inputs include the args.me corpus and top 50 topics on which uploaded retrieval models will be judged. Entries are ranked by nDCG@5 but TIRA also returns nDCG@10, nDCG, and QrelCoverage@10
Figure 1 represent the distribution of Combined Quality in test set. The following tables show the distribution of predicted values using different evaluation approaches.

After retrieving the top 50 documents using the baseline, retrieved documents will then be re-ranked using our best possible model combining both POS bi- and unigrams

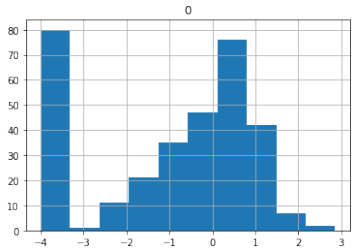

Figure 1: Distribution of qualities in test set

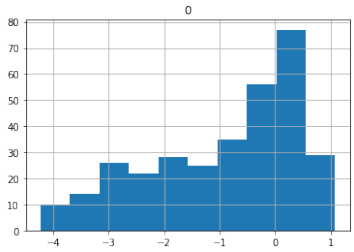

Figure 2: GloVe and POS bigrams predictions

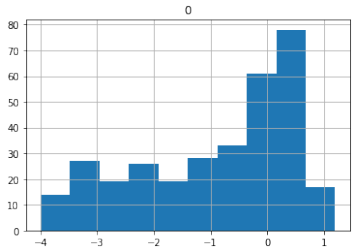

Figure 3: BERT embeddings predictions

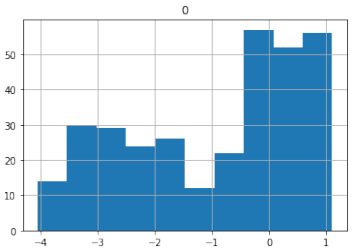

Figure 4: POS bi- and unigrams predictions

The nDCG of the baseline and reranking approach using the regression model.

Model	nDCG@5	nDCG@10	nDCG	QrelCoverage@10
Baseline: TF-IDF	-	**0.5046**	-	-
Re-Ranked	-	0.1764	-	-

4. Conclusion and Proposals

This work shows our attempt at creating a retrieval model using tf-idf methodology and re-ranking the top 50 retrieved documents using four different approaches. The biggest problem for us and our approaches was that we are unable to predict when premise is not an argument at all, so having -4 rating as combined quality. As Figure 1 shows, corpus possess a lot of premises which are not tagged as argument and our approaches predicting at most 15 of them as non argument while test set has 80 of them. We were not able to get better results than our baseline and there is still a lot of space to improve the used approaches. For example, trying to use different types of POS tags creating embeddings or scale up dimensions in embeddings created by GloVe at least to 300. As well as natural language prepossessing of the corpus could be used or hyperpamater fine-tuning of models. Or simply trying different methods that take the original ranking into account.

References

[1] C. M. Jeffrey Pennington, Richard Socher, GloVe: Global Vectors for Word Representation, 2014.

[2] K. L. K. T. Jacob Devlin, Ming-Wei Chang, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018.

[3] J. K. M. P. M. H. B. S. Yamen Ajjour, Henning Wachsmuth, args.me corpus, 2019.

[4] M. H. M. P. Lukas Gienapp, Benno Stein, Webis Argument Quality Corpus 2020 (Webis-ArgQuality-20), 2020.

[5] E. L. Steven Bird, NLTK: The Natural Language Toolkit, 2004.

[6] C. G. Tianqi Chen, XGBoost: A Scalable Tree Boosting System, 2016.