Approach: Submodular Inference

A straightforward approach to citation recommendation could be:

- Turn the documents into some numerical representation, e.g. TF-IDF.
- Compute the similarity between the query and each candidate document using a similarity function, e.g. the cosine similarity.
- Recommend the top-K most similar documents.

Risk? → redundancy!!! All the recommended citations may look the same.

When selecting the citations, one should balance similarity to the query and diversity of the recommended citations.

A scoring function that balances these two properties is typically submodular, and finding the citation list that maximizes it is NP-hard. However, submodular functions enjoy a key property: selecting the citations one by one with a simple, greedy algorithm is near-optimal.*

Approach: Document Representation

To represent the document, we use Sentence-BERT [Reimers & Gurevych EMNLP 2019]: a neural approach to embed a whole sentence/paragraph/short document into a vector using any pre-trained BERT model. We fine-tune it in a supervised manner:

- Annoted training set: the documents and their citations in the corpus (the citation graph).
- Positive examples for query q: all documents d_i’s with number of nodes in the shortest citation path from q ≤ 3.
- Negative examples: all the others. To limit the training time, we only use subsets of the negative examples, selected with three different strategies: Random, Nearest (to the query in similarity) and Farthest.

As fine-tuning objectives, we use:

- Twin aka “Siamese” network-style: given a query q, and a positive or negative candidate, d, we minimize the mean squared difference between their predicted and target similarities.
- Triplet network-style: given the query q, a positive candidate, d^+ and a negative candidate, d^−, we impose that the predicted similarity s(q,d^+) be larger than s(q,d^−) by a margin: triplet loss = max[s(q,d^+) − s(q,d^−) + 1, 0].

Experiments and Results

**Dataset:** ACL Anthology Network corpus (AAN) [4]: a dataset of 22,085 papers in the field of computational linguistics. Papers + meta-information. Same experimental setup as [2] excluding papers with no references and using the standard training (16,128 papers from 1900 to 2010), dev/validation (1,060 papers from 2011) and test (1,161 papers from 2012) splits.

**Performance evaluation:** Mean Reciprocal Rank (MRR) and F1@k score

**Compared approaches:**

- ElasticSearch with Okapi BM25
- Citeomatic [3]
- Our previous submodular approach, SubRef [2]
- The proposed method with a simple top-K inference
- The proposed method with submodular inference

Main results:

<table>
<thead>
<tr>
<th>Method</th>
<th>MRR</th>
<th>F1@10</th>
<th>F1@20</th>
<th>F1@50</th>
<th>F1@100</th>
</tr>
</thead>
<tbody>
<tr>
<td>ElasticSearch</td>
<td>0.2437</td>
<td>0.0071</td>
<td>0.0063</td>
<td>0.0046</td>
<td>0.0032</td>
</tr>
<tr>
<td>Citeomatic</td>
<td>0.3085</td>
<td>0.1281</td>
<td>0.1339</td>
<td>0.0940</td>
<td>0.0548</td>
</tr>
<tr>
<td>Select</td>
<td>0.3777</td>
<td>0.1590</td>
<td>0.1472</td>
<td>0.0059</td>
<td>0.0049</td>
</tr>
<tr>
<td>Select + Rank</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SubRef (best on dev)</td>
<td>0.3320</td>
<td>0.1310</td>
<td>0.1264</td>
<td>0.0011</td>
<td>0.0021</td>
</tr>
<tr>
<td>Siamese + BM25</td>
<td>0.3320</td>
<td>0.1310</td>
<td>0.1264</td>
<td>0.0011</td>
<td>0.0021</td>
</tr>
<tr>
<td>Siamese + QAIv2</td>
<td>0.4431</td>
<td>0.1978</td>
<td>0.1839</td>
<td>0.1327</td>
<td>0.0918</td>
</tr>
</tbody>
</table>

The proposed approach has outperformed all the compared approaches in all metrics on this dataset. We will extend our evaluation to other datasets in the near future.