<table>
<tr><th rowspan="2">cases</th><th>doc_1</th><th>doc_2</th><th>decision</th><th>id</th></tr>
<tr><td>
<table>
<tr><th>authors</th><td>Y. Liang
Rui Cao
Jie Zheng
Jie Ren
Ling Gao</td></tr>
<tr><th>title</th><td>Learning to Remove: Towards Isotropic Pre-trained BERT Embedding</td></tr>
<tr><th>publication_date</th><td>2021-04-12 00:00:00</td></tr>
<tr><th>source</th><td>SupportedSources.SEMANTIC_SCHOLAR</td></tr>
<tr><th>journal</th><td></td></tr>
<tr><th>volume</th><td></td></tr>
<tr><th>doi</th><td>10.1007/978-3-030-86383-8_36</td></tr>
<tr><th>urls</th><td>https://www.semanticscholar.org/paper/ab151c1ca0479b677003ef200018b93e983aa0ec</td></tr>
<tr><th>id</th><td>id-9088774381427119780</td></tr>
<tr><th>abstract</th><td>None</td></tr>
<tr><th>versions</th><td></td></tr>
</table>
</td><td>
<table>
<tr><th>authors</th><td>Yuxin Liang
Rui Cao
Jie Zheng
Jie Ren
Ling Gao</td></tr>
<tr><th>title</th><td>Learning to Remove: Towards Isotropic Pre-trained BERT Embedding</td></tr>
<tr><th>publication_date</th><td>2021-08-27 00:00:00</td></tr>
<tr><th>source</th><td>SupportedSources.INTERNET_ARCHIVE</td></tr>
<tr><th>journal</th><td></td></tr>
<tr><th>volume</th><td></td></tr>
<tr><th>doi</th><td></td></tr>
<tr><th>urls</th><td>https://web.archive.org/web/20210902205317/https://arxiv.org/pdf/2104.05274v2.pdf</td></tr>
<tr><th>id</th><td>id8517331927159074556</td></tr>
<tr><th>abstract</th><td>Pre-trained language models such as BERT have become a more common choice of natural language processing (NLP) tasks. Research in word representation shows that isotropic embeddings can significantly improve performance on downstream tasks. However, we measure and analyze the geometry of pre-trained BERT embedding and find that it is far from isotropic. We find that the word vectors are not centered around the origin, and the average cosine similarity between two random words is much higher than zero, which indicates that the word vectors are distributed in a narrow cone and deteriorate the representation capacity of word embedding. We propose a simple, and yet effective method to fix this problem: remove several dominant directions of BERT embedding with a set of learnable weights. We train the weights on word similarity tasks and show that processed embedding is more isotropic. Our method is evaluated on three standardized tasks: word similarity, word analogy, and semantic textual similarity. In all tasks, the word embedding processed by our method consistently outperforms the original embedding (with average improvement of 13% on word analogy and 16% on semantic textual similarity) and two baseline methods. Our method is also proven to be more robust to changes of hyperparameter.</td></tr>
<tr><th>versions</th><td></td></tr>
</table>
</td><td>DUPLICATES</td><td>221</td></tr>
</table>