	Dataset	F-score	
		Original	Ours
Tf-Idf	DUC	25.4	25.7
TextRank	*Inspec*	36.2	10.0
SingleRank	DUC	27.2	24.9
ExpandRank	DUC	31.7	26.4
KeyCluster	*Inspec*	43.6	38.9

Table 3: Original vs. re-implementation scores

of TextRank[3], and are confident that our implementation is correct. It is also worth mentioning that using our re-implementation of SingleRank, we are able to match the best scores reported by Mihalcea and Tarau (2004) on *Inspec*.

We score 2 and 5 points less than Wan and Xiao's (2008) implementations of SingleRank and ExpandRank, respectively. We speculate that document pre-processing (e.g., stemming) has contributed to the discrepancy, but additional experiments are needed to determine the reason.

SingleRank vs. TextRank Figure 1 shows that SingleRank behaves very differently from TextRank. As mentioned in Section 3.2.3, the two algorithms differ in three major aspects. To determine which aspect is chiefly responsible for the large difference in their performance, we conduct three "ablation" experiments. Each experiment modifies exactly one of these aspects in SingleRank so that it behaves like TextRank, effectively ensuring that the two algorithms differ only in the remaining two aspects. More specifically, in the three experiments, we (1) change SingleRank's window size to 2, (2) build an unweighted graph for SingleRank, and (3) incorporate TextRank's way of forming keyphrases into SingleRank, respectively. Figure 2 shows the resultant curves along with the SingleRank and TextRank curves on *Inspec* taken from Figure 1b. As we can see, the way of forming phrases, rather than the window size or the weight assignment method, has the largest impact on the scores. In fact, after incorporating TextRank's way of forming phrases, SingleRank exhibits a remarkable drop in performance, yielding a curve that resembles the TextRank curve. Also note that SingleRank achieves better recall values than TextRank. To see the reason, recall that TextRank requires that every word of a gold keyphrase must appear among the top-ranked unigrams. This is a fairly strict criterion, especially in comparison to SingleRank, which does not require all unigrams of a gold keyphrase to be present in the top-ranked list. We observe similar trends for the other datasets.

[3] http://github.com/sharethis/textrank

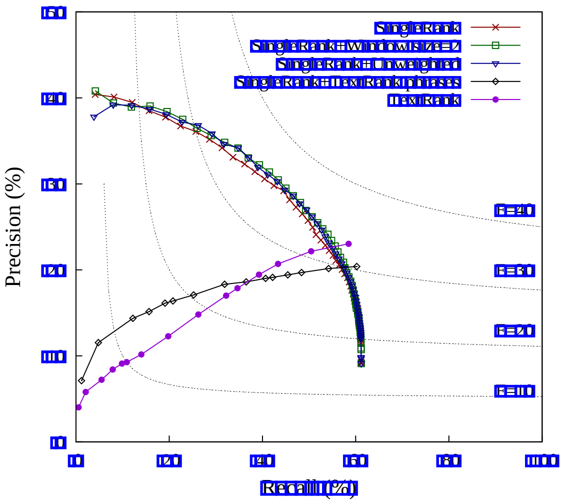

Figure 2: Ablation results for SingleRank on *Inspec*

5 Conclusions

We have conducted a systematic evaluation of five state-of-the-art unsupervised keyphrase extraction algorithms on datasets from four different domains. Several conclusions can be drawn from our experimental results. First, to fully understand the strengths and weaknesses of a keyphrase extractor, it is essential to evaluate it on multiple datasets. In particular, evaluating it on a single dataset has proven inadequate, as good performance can sometimes be achieved due to certain statistical characteristics of a dataset. Second, as demonstrated by our experiments with TextRank and SingleRank, post-processing steps such as the way of forming keyphrases can have a large impact on the performance of a keyphrase extractor. Hence, it may be worthwhile to investigate alternative methods for extracting candidate keyphrases (e.g., Kumar and Srinathan (2008), You et al. (2009)). Finally, despite the large amount of recent work on unsupervised keyphrase extraction, our results indicated that Tf-Idf remains a strong baseline, offering very robust performance across different datasets.