	DUC		*Inspec*		NUS		ICSI	
	Parameter	F	Parameter	F	Parameter	F	Parameter	F
Tf-Idf	N = 14	27.0	N = 14	36.3	N = 60	6.6	N = 9	12.1
TextRank	T = 100%	9.7	T = 100%	33.0	T = 5%	3.2	T = 25%	2.7
SingleRank	N = 16	25.6	N = 15	35.3	N = 190	3.8	N = 50	4.4
ExpandRank	N = 13	26.9	N = 15	35.3	N = 177	3.8	N = 51	4.3
KeyCluster	$m = 0.9n$	14.0	$m = 0.9n$	40.6	$m = 0.25n$	1.7	$m = 0.9n$	3.2

Table 2: Best parameter settings. N is the number of predicted keyphrases, T is the percentage of vertices selected as keywords in TextRank, m is the number of clusters in KeyCluster, expressed in terms of n, the fraction of candidate words.

ExpandRank Consistent with Wan and Xiao (2008), ExpandRank beats SingleRank on DUC when a small number of phrases are predicted, but their difference diminishes as more phrases are predicted. On the other hand, their performance is indistinguishable from each other on the other three datasets. A natural question is: why does ExpandRank improve over SingleRank only for DUC but not for the other datasets? To answer this question, we look at the DUC articles and find that in many cases, the 5-nearest neighbors of a document are on the same topic involving the same entities as the document itself, presumably because many of these news articles are simply updated versions of an evolving event. Consequently, the graph built from the neighboring documents is helpful for predicting the keyphrases of the given document. Such topic-wise similarity among the nearest documents does not exist in the other datasets, however.

KeyCluster As in TextRank, KeyCluster does not always yield a drop in precision as recall improves. This, again, may be attributed to the fact that the KeyCluster curves are generated by varying the number of clusters rather than the number of predicted keyphrases, as well as the way keyphrases are formed from the exemplars. Another reason is that the frequent Wikipedia unigrams are excluded during post-processing, making KeyCluster more resistant to precision drops. Overall, KeyCluster performs slightly better than TextRank on DUC and ICSI, yields the worst performance on NUS, and scores the best on *Inspec* when the number of clusters is high. These results seem to suggest that KeyCluster works better if more clusters are used.

Best parameter settings Table 2 shows for each system the parameter values yielding the best F-score on each dataset. Two points deserve mention. First, in comparison to SingleRank and ExpandRank, Tf-Idf outputs fewer keyphrases to achieve its best F-score on most datasets. Second, the systems output more keyphrases on NUS than on other datasets to achieve their best F-scores (e.g., 60 for Tf-Idf, 190 for SingleRank, and 177 for ExpandRank). This can be attributed in part to the fact that the F-scores on NUS are low for all the systems and exhibit only slight changes as we output more phrases.

Our re-implementations Do our duplicated systems yield scores that match the original scores? Table 3 sheds light on this question.

First, consider KeyCluster, where our score lags behind the original score by approximately 5%. An examination of Liu et al.'s (2009b) results reveals a subtle caveat in keyphrase extraction evaluations. In *Inspec*, not all gold-standard keyphrases appear in their associated document, and as a result, none of the five systems we consider in this paper can achieve a recall of 100. While Mihalcea and Tarau (2004) and our reimplementations use *all* of these gold-standard keyphrases in our evaluation, Hulth (2003) and Liu et al. address this issue by using as gold-standard keyphrases only those that appear in the corresponding document when computing recall.[2] This explains why our KeyCluster score (38.9) is lower than the original score (43.6). If we follow Liu et al.'s way of computing recall, our reimplementation score goes up to 42.4, which lags behind their score by only 1.2.

Next, consider TextRank, where our score lags behind Mihalcea and Tarau's original score by more than 25 points. We verified our implementation against a publicly available implementation

[2] As a result, Liu et al. and Mihalcea and Tarau's scores are not directly comparable, but Liu et al. did not point this out while comparing scores in their paper.