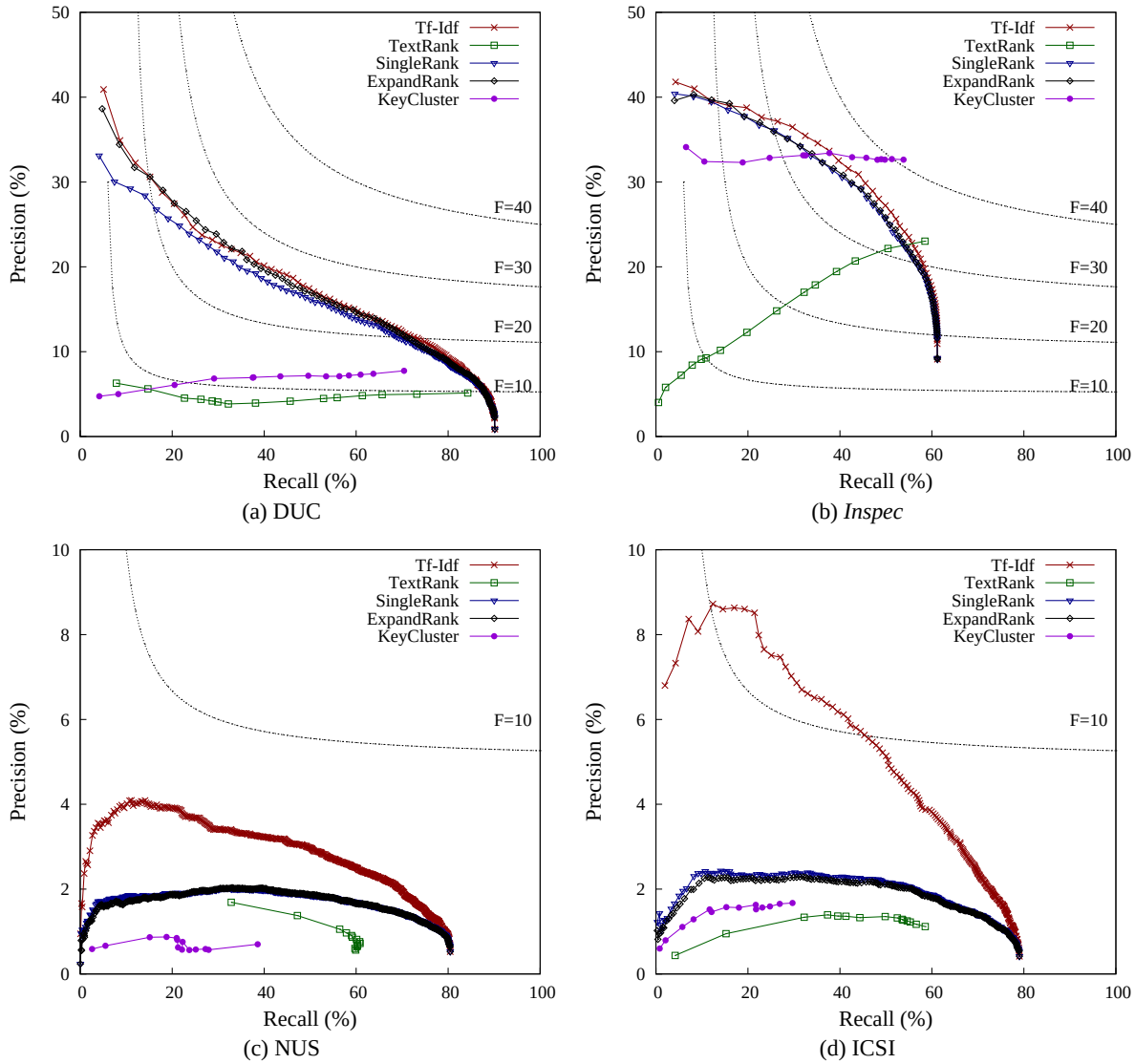

Figure 1: Precision-recall curves for all four datasets

"advanced" and "learning" are now combined to form one keyphrase (and hence the number of predicted keyphrases decreases). In other words, it is possible to see a simultaneous rise in precision and recall in a TextRank curve. A natural question is: why does is happen only for *Inspec* but not the other datasets? The reason could be attributed to the fact that *Inspec* is composed of abstracts: since the number of keyphrases that can be generated from these short documents is relatively small, precision may not drop as severely as with the other datasets even when all of the unigrams are used to form keyphrases.

On average, TextRank performs much worse compared to Tf-Idf. The curves also prove TextRank's sensitivity to T on *Inspec*, but not on the other datasets. This certainly gives more insight into TextRank since it was evaluated on *Inspec* only for T=33% by Mihalcea and Tarau (2004).

SingleRank SingleRank, which is supposed to be a simple variant of TextRank, surprisingly exhibits very different performance. First, it shows a more intuitive nature: precision drops as recall increases. Second, SingleRank outperforms TextRank by big margins on all the datasets. Later, we will examine which of the differences between them is responsible for the differing performance.